THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *28 juin 2016* **par :**

MARWA THLITHI

Segmentation et regroupement en chanteurs.
Application aux enregistrements ethnomusicologiques.

JURY

G. Linarès	Rapporteur	LIA
S. Meignier	Rapporteur	LIUM
M. Desainte-Catherine	Présidente	LaBRI
R. André-Obrecht	Directrice	IRIT
J. Pinquier	Co-directeur	IRIT
T. Pellegrini	Co-encadrant	IRIT

École doctorale et spécialité :
MITT : Image, Information, Hypermédia
Unité de Recherche :
Institut de Recherche en Informatique de Toulouse (UMR 5505)
Directeur(s) de Thèse :
Régine André-Obrecht et *Julien Pinquier*
Rapporteurs :
Sylvain Meignier et *Georges Linarès*

Remerciements

Je tiens tout d'abord à adresser mes sincères remerciements aux encadrants de ma thèse : Régine André-Obrecht, Julien Pinquier et Thomas Pellegrini qui ont su m'accorder leur confiance et me guider tout au long de ma thèse. Ils m'ont beaucoup appris et ils m'ont beaucoup sacrifié de leurs temps.

Je remercie ma directrice de thèse Régine pour l'accompagnement de mes travaux ainsi que pour ses précieux conseils. Je remercie également mon co-directeur de thèse Julien pour la qualité de l'encadrement, le suivi régulier de mes tâches, ses idées et son assistance sur les sujets administratifs. Je remercie Thomas encadrant de ma thèse pour sa disponibilité ainsi que ses contributions efficaces qui m'ont beaucoup aidé à avancer. C'est grâce à eux que ce travail était possible et réalisé.

Je suis également très reconnaissante aux Messieurs Sylvain Meignier et Georges Linarès d'avoir accepté d'être les rapporteurs de ce travail de thèse, et à Madame Myriam Desainte-Catherine d'avoir accepté de participer à mon jury de soutenance.

Je remercie tous les membres de l'équipe SAMoVA pour leurs conseils et leurs sympathies et en particulier mes collègues de bureau : Bouchra Soukkarieh, François-Xavier Decroix et Philippe Ercolessi. Je remercie également Maxime Le Coz et Patrice Guyot pour leurs conseils et leurs aides.

Je tiens aussi à remercier tous les membres du LIUM qui m'ont offert l'opportunité d'être parmi eux dans le cadre d'un poste d'ATER pendant ma dernière année de thèse. Je les remercie pour l'accueil chaleureux et leurs aides qui m'ont permis de mener ces travaux jusqu'au bout.

J'adresse ma gratitude à mes amis : Bouchra, Farouk, Sameh, Monia, Mathieu, Sahar, Abdessalam, Rajoua et bien d'autres.

Un très grand merci à mon fiancé Alaeddine qui s'est déplacé pour assister à ma soutenance. Encore merci à lui qui m'a supporté, qui m'a toujours encouragé et qui a cru en mes capacités à porter et mener à bien ce projet.

Un très grand merci à ma famille qui m'a soutenu et supporté tout au long de la thèse, avec une pensée particulière à mon père et ma mère. L'amour et la confiance qu'ils m'ont toujours témoignés m'ont permis de traverser ces années de thèse et tous les moments difficiles de la vie.

A mon père et ma mère

Résumé

Cette thèse est réalisée dans le cadre du projet ANR CONTINT DIADEMS sur l'indexation de documents ethnomusicologiques sonores. Les données que nous traitons sont fournies par les partenaires ethnomusicologues du projet et elles sont issues des archives du Musée de l'Homme de Paris. Les travaux effectués lors de cette thèse consistent à développer des méthodes permettant de faire une structuration automatique des documents musicaux et ethnomusicologiques basée sur les personnes.

Cette thèse aborde le sujet encore inexploré à notre connaissance de la segmentation et du regroupement en chanteurs dans des enregistrements musicaux. Nous proposons un système complet pour ce sujet en s'inspirant des travaux réalisés en segmentation et regroupement en locuteurs. Ce système est conçu pour fonctionner aussi bien sur des enregistrements musicaux de type studio que sur des enregistrements musicaux réalisés dans des conditions terrain. Il permet, tout d'abord, de découper les zones de chant en des segments acoustiquement homogènes, i.e. en groupe de chanteur(s) afin d'avoir une segmentation en tours de chant. Ensuite, une phase de regroupement est effectuée afin de rassembler tous les segments chantés par un même groupe de chanteur(s) dans une seule classe.

Notre première contribution est la définition de la notion de « tour de chant » et la proposition de règles d'annotation manuelle d'un enregistrement en des segments de tours de chant. La deuxième est la proposition d'une méthode de paramétrisation de la voix des chanteurs en implémentant une stratégie de sélection de bandes fréquentielles pertinentes basée sur la variance de celles-ci. La troisième est l'implémentation d'un algorithme de segmentation dynamique adapté à un contexte de chant en utilisant le Critère d'Information Bayésien (BIC). La quatrième est la proposition d'une méthode de Décision par Consolidation *A Posteriori*, nommée DCAP, pour pallier au problème de variabilité du paramètre de pénalité du BIC. En effet, comme le choix *a priori* d'une valeur optimale de ce paramètre n'est pas possible, nous effectuons un vote majoritaire sur plusieurs sorties de segmentations obtenues avec différentes valeurs de ce paramètre. Des gains d'environ 8% et 15% sont obtenus sur nos deux corpus avec cette méthode par rapport à une valeur standard du paramètre de pénalité. La cinquième est l'adaptation de la méthode DCAP pour la réalisation de l'étape de regroupement en chanteurs.

Abstract

This work was done in the context of the ANR CONTINT DIADEMS project on indexing ethno-musicological audio recordings. The data that we are studying are provided by the *Musée de l'Homme*, Paris, within the context of this project. The work performed in this thesis consists of developing automatic structuring methods of musical and ethno-musicological documents based on the persons.

This thesis touchs on an unexplored subject in our knowledge of the segmentation and clustering in singers of musical recordings. We propose a complete system in this subject that we called *singer diarization* by analogy with *speaker diarization* system on speech context. Indeed, this system is inspired from existing studies performed in *speaker diarization* and is designed to work on studio music recordings as well as on recordings with a variable sound quality (done outdoors). The first step of this system is the segmentation in singer turns which consists of segmenting musical recordings into segments "acoustically homogeneous" by singer group. The second step is the clustering which consists of labelling all segments produced by the same group of singers with a unique identifier.

Our first contribution involved the definition of the term « singer turns » and the proposal of rules for manual annotation in singer turns segments. The second consisted in the proposal of a feature extraction method for the characterization of singer voices by implementing a method to select the frequency coefficients, which are the most relevant, based on the variance of these coefficients. The third is the implementation of a dynamic segmentation algorithm adapted to the singing context by using the Bayesian Information Criterion (BIC). The fourth is the proposal of a method, called DCAP, to take *a posteriori* decisions in order to avoid the variability problem of the BIC penalty parameter. Indeed, *a priori* choice of an optimal value for this parameter is not possible. This led us to perform a majority voting on a several segmentations obtained with different values of this parameter. A gain of about 8% and 15% is obtained on our two corpora with this method compared to the results found with a standard value of the penalty parameter. The fifth is the adaptation of our DCAP method in order to perform singer clustering step.

Table des matières

Tables des figures

Liste des tableaux

Liste des acronymes

BIC Bayesian Information Criterion

CE Cross Entropy

CLR Cross Likelihood Ratio

DCAP Décision par Consolidation *A posteriori*

DER Diarization Error Rate

EM Expectation Maximization

EVSM Eigen Vector Space Model

FBANK Filter Bank

GLR Generalized Likelihood Ratio

GMM Gaussian Mixture Models

HMM Hidden Markov Models

ILP Integer Linear Programming

KL Kullback-Leibler

KL2 Symetric Kullback-Leibler

MAP Maximum *A Posteriori*

MFCC Mel Frequency Cepstral Coefficient

MPH Multi-Probe Histogram

NIST National Institute of Standards and Technology

PCA Principal Component Analysis

RASTA-PLP RelAtiveSpecTrAl-Perceptual Linear Prediction

PLP Perceptual Linear Prediction

RDCAP Regroupement avec Décision par Consolidation *A Posteriori*

SRL Segmentation et Regroupement en Locuteurs

SVM Support Vector Machines

UBM Universel Background Model

ZCR Zero Crossing Rate

Chapitre 1

Introduction

1.1. Contexte

Dans un contexte de transformation et de développement numérique fulgurant, l'idée d'indexation automatique des données multimédia est née dans un objectif principal d'y faciliter l'accès. En effet, en observant la quantité des données audiovisuelles sur le Web ou à travers les médias traditionnels tels que la radio et la télévision, nous présumons que le besoin consistant à rechercher et créer des méthodes d'indexation de ces contenus, est bien réel.

En plus de l'explosion des données numériques, l'être humain a toujours tendance à hiérarchiser et structurer l'information perçue. Un document sonore peut être structuré automatiquement de bien des manières en fonction de l'objectif final. L'indexation automatique vise à structurer le flux des données en se basant sur des descripteurs de haut niveau ou de bas niveau. S'il s'agit d'indexation d'un fond de documents par exemple, nous nous poserons vraisemblablement les questions : sommes-nous en présence de parole ? De musique ? A quels moments ? Qui parle ? Qui chante ? Etc. Structuration et indexation sont alors très liées et l'une des premières étapes d'inférence de cette structure est de découper puis d'étiqueter des zones ou segments dits « acoustiquement homogènes ». Cette étape consiste à extraire des segments homogènes, ne contenant que les données qui répondent à une condition acoustique. Les conditions acoustiques existent sous différentes formes telles que la parole, la musique, le chant, le bruit, le silence, la voix d'homme, la voix de femme, le signal en bande large ou en bande étroite. Nous pouvons citer, à titre d'exemple, quelques études faites dans ce domaine portant sur la recherche des composantes primaires d'un document sonore comme la parole, la musique, le chant ou les sons clés, la segmentation en des zones homogènes en phonèmes et la segmentation en tours de parole (zones homogènes par locuteur).

Plusieurs travaux réalisés en traitement automatique de la parole utilisent la segmentation en des zones homogènes comme une étape de prétraitement indispensable. Parmi eux figurent les travaux en détection de la parole, en regroupement en locuteurs, en reconnaissance des locuteurs (vérification, identification), etc. Ces travaux sont bien avancés et ont donné de très bons résultats. Des études en traitement automatique de la musique ont été menées telles que la détection de la musique, la reconnaissance des instruments ou la segmentation en structures musicales. Comme la plupart des systèmes de détection de musique classe le chant dans la catégorie *musique*, les recherches en détection de musique peuvent servir comme prétraitement pour d'autres tâches de traitement du chant telles que la détection du chant, l'identification des chanteurs ou le suivi des chanteurs. Ces tâches font partie des nombreux

travaux effectués en traitement automatique du chant. Bien que plusieurs recherches ont été menées ces dernières années en traitement de chant, aucune n'a abordé le sujet de la segmentation en tours de chant, alors que cette tâche peut être utilisée comme étape de prétraitement indispensable permettant de faire par la suite de la détection, de la reconnaissance ou de la vérification d'un chanteur.

Dans ce cadre, l'équipe SAMoVA (**S**tructuration, **A**nalyse et **Mo**délisation des documents **V**idéo et **A**udio), formée en 2002 par des acteurs issus du monde de la parole et de la vidéo, travaille essentiellement sur l'indexation par le contenu des documents multimédia. L'équipe a principalement commencé par réaliser des travaux en indexation par le contenu des documents de parole. Ensuite, les travaux se sont étendus pour traiter d'autres types de sons comme la voix chantée et sa détection en tenant compte du contexte polyphonique.

Dans cette même optique d'ouverture de l'équipe sur le domaine du traitement de la musique et de la voix chantée, elle s'est engagée comme porteur du projet ANR DIADEMS[1]. Ce projet a comme objectif d'analyser les enregistrements ethnomusicologiques issus des archives du Laboratoire d'Ethnologie et de Sociologie Comparative (LESC) datant du début de XX^{e} siècle jusqu'à nos jours. Le corpus de ce projet est riche de plusieurs types de sons (enregistrements musicaux, interviews, contes, etc.) provenant de nombreuses ethnies différentes dans le monde. Il possède aussi la spécificité d'être composé d'enregistrements effectués sur le terrain dans des conditions rarement optimales. L'indexation de telles données sonores lève de nouveaux défis de recherche. Dans le but de faciliter l'accès aux enregistrements de ce corpus, le projet DIADEMS vise à fournir des outils d'indexation d'un document sonore ou d'une collection de documents comme la classification des instruments, la détection de la musique, la détection de la voix chantée, la segmentation en tours de parole et la segmentation en tours de chant, en limitant au maximum les phases d'apprentissage.

1.2. Problématique

Dans ce contexte d'indexation et d'étude du chant, nous nous sommes intéressés aux zones de chant dans un document musical. Parmi les questions qui peuvent être posées en présence de chant et en faisant une structuration basée sur les personnes, est « **qui chante et quand ?** ». En présence de parole, cette question s'est également posée en s'intéressant à « **qui parle et quand ?** ». La communauté du traitement automatique de la parole a beaucoup étudié cette problématique et ainsi plusieurs versions de système de segmentation et regroupement en locuteurs ont été mises en place. Ces systèmes consistent à réaliser, tout d'abord, une segmentation en tours de parole afin de découper le flux de la parole en des segments « acoustiquement homogènes » par locuteurs ; par la suite un regroupement en locuteurs est effectué pour regrouper les segments de parole prononcés par un même locuteur, dans une seule et unique classe.

[1] https://www.irit.fr/recherches/SAMOVA/DIADEMS/fr/accueil/

Comme la parole et le chant sont tous les deux produits par l'humain, pour répondre à la question « **qui chante et quand ?** », nous nous sommes beaucoup inspirés des travaux réalisés sur la parole. La problématique de segmentation et regroupement en chanteurs est illustrée dans la Figure 1. 1.

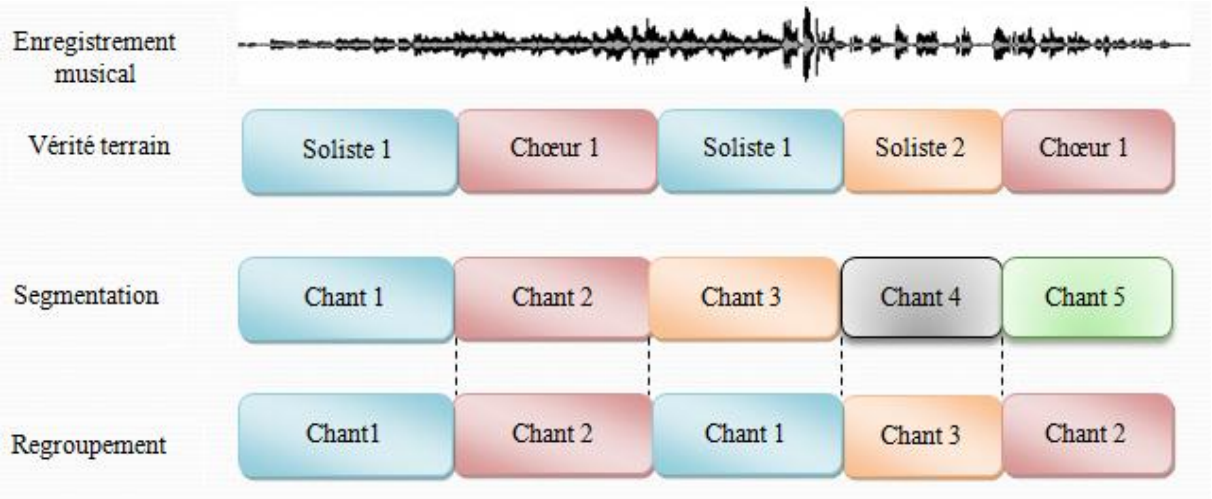

Figure 1. 1 – Illustration de la problématique de segmentation et regroupement en chanteurs.

A l'instar de la plupart des systèmes de segmentation et regroupement en locuteurs, la réalisation d'un système de segmentation et regroupement en chanteurs implique l'existence de deux étapes :

- la segmentation en tours de chant. Celle-ci consiste à découper le flux contenant du chant en des segments « acoustiquement homogènes », c'est-à-dire ici à détecter le changement de groupe de chanteur(s),

- le regroupement en chanteurs. Cette étape permet de rassembler les segments chantés par le même groupe de chanteur(s), ce qui revient à identifier ces segments par une même étiquette.

Pour le chant, nous utilisons la notion de groupe de chanteur(s) car nous avons souvent plusieurs chanteurs qui chantent en même temps. Ainsi, un groupe de chanteur(s) peut être constitué d'un seul chanteur ou de plusieurs. Hors en parole, cette notion n'est pas utilisée car les groupes de locuteurs (parole superposée) sont beaucoup plus rares. En effet, pour des raisons de politesse et d'intelligibilité, les locuteurs ne parlent pas (ou de manière très brève) les uns sur les autres.

De manière générale, les objectifs de cette thèse sont de proposer des outils de segmentation en tours de chant et de regroupement en chanteurs en proposant des méthodes génériques qui conduisent à des performances acceptables aussi bien sur des enregistrements

réalisés en studio, que sur des enregistrements effectués sur le « terrain », comme ceux présents dans le projet DIADEMS.

Par analogie avec la segmentation en tours de parole, dans notre étude en segmentation en tours de chant, nous nous sommes fondés sur une approche de type probabiliste couramment utilisée pour la segmentation audio et en particulier pour la parole : le Critère d'Information Bayésien (BIC). Nos premières études en segmentation avec ce critère ont montré l'importance de deux paramètres de l'algorithme, à savoir la taille de la fenêtre d'analyse et le facteur de pénalité. En essayant de les ajuster, nous nous sommes aperçus que leurs valeurs variaient énormément d'un enregistrement à l'autre, alors que ce n'est pas le cas pour la parole. Cette observation nous a conduit à examiner la possibilité de développer un système de segmentation utilisant une taille de fenêtre d'analyse dynamique et évitant tout choix *a priori* du facteur de pénalité. Nos études avancées en segmentation en tours de chant ont relevé un problème de caractérisation de la voix chantée. En effet, les paramètres classiques utilisés en parole (MFCC) se sont avérés non adaptés à notre contexte de segmentation en des zones homogènes par groupe de chanteur(s). Cela nous a conduit à étudier d'autres méthodes de paramétrisation de la voix chantée.

Afin de regrouper en chanteurs, nous avons là aussi procédé par analogie avec le regroupement en locuteurs, où la plupart des systèmes réalisés est aussi fondée sur le BIC. Pour cette étape, nous avons eu aussi des difficultés pour fixer la valeur du facteur de pénalité. Ce problème nous a amené à examiner la possibilité d'éviter le choix *a priori* de ce facteur.

Les contributions majeures des recherches présentées dans ce document sont :
- la proposition d'une méthode de paramétrisation de la voix des chanteurs,
- l'implémentation d'un algorithme de segmentation dynamique, adapté au contexte du chant,
- la proposition d'une méthode de fusion évitant le choix *a priori* du facteur de pénalité du BIC,
- une adaptation de la méthode précédente pour la réalisation du regroupement en chanteurs.

1.3. Organisation du manuscrit

La problématique et les objectifs de cette thèse nécessitent une organisation du manuscrit en cinq chapitres.

Afin de présenter les travaux qui s'inscrivent dans le contexte de cette recherche, nous commençons, dans le chapitre 2, par une brève revue de l'état de l'art sur les recherches en traitement automatique de chant et en segmentation et regroupement en locuteurs. Nous nous attardons en particulier sur les outils développés pour en décrire divers aspects : séparation monophonie / polyphonie, détection de chant, segmentation solo / chœur, segmentation en structure musicale, segmentation en tours de parole et regroupement en locuteurs.

Le chapitre 3 est dédié à la définition des principaux concepts de ce travail, aux présentations du corpus de travail et du guide d'annotation que nous avons mis en place ainsi qu'à la description des métriques que nous utilisons pour évaluer nos systèmes de segmentation et de regroupement.

Les chapitres 4 et 5 sont consacrés aux deux étapes que nous avons décrites : d'une part la segmentation en tours de chant, d'autre part le regroupement en chanteurs. Pour chacun de ces deux thèmes, nous présentons, au sein de chaque chapitre, notre approche et l'ensemble des expériences que nous avons menées sur un corpus d'enregistrements studio afin de valider celle-ci.

Enfin, dans le chapitre 6, nous présentons le corpus du projet DIADEMS ainsi que ses particularités. Ensuite, nous illustrons les expérimentations issues de l'application de notre méthode de segmentation en tours de chant et de notre approche regroupement en chanteurs, qui ont été validées dans les chapitres 4 et 5, sur les données du corpus studio. Des étapes de prétraitement sont proposées, ce qui induit la mise en place d'une chaîne de traitement complète pour les données du corpus DIADEMS. De plus, des améliorations sont proposées pour l'étape de regroupement en chanteurs sur ce corpus.

Chapitre 2

Etat de l'art

Sommaire

2.1. Introduction – Chant, segmentation et regroupement en locuteurs

La tâche de segmentation et regroupement en chanteurs n'est pas un sujet de recherche très exploré, alors que pour la parole, la tâche de segmentation et regroupement en locuteurs est largement étudiée, les systèmes récents donnant des performances plus qu'honorables. Ce chapitre a pour objectif de montrer, de manière certainement partielle, ce qui est connu des travaux dans ce domaine, sur le chant et sur la parole, afin de pointer les différences potentielles et de se positionner par rapport à l'existant.

Du fait de la jeunesse de ce sujet d'études, il est nécessaire de préciser la notion de chant elle-même, et les caractéristiques de ce type de son, ce qui fait l'objet du paragraphe 2.2. Ce positionnement est complété, dans le paragraphe 2.3, d'une étude sur quelques systèmes en traitement automatique du chant ainsi qu'une description des paramètres acoustiques utilisés dans les méthodes existantes.

En partant de l'idée que le chant est produit par la voix humaine comme la parole, nous avons décidé de faire un état de l'art sur les systèmes de segmentation et regroupement en locuteurs ainsi que sur les différents paramètres et approches utilisés dans ces systèmes afin de s'en inspirer et trouver une méthode pour réaliser la même tâche sur le chant. Au cours du paragraphe 2.4 consacré à cette tâche de segmentation et regroupement en locuteurs, nous commençons par définir les méthodes de paramétrisation existantes pour cette tâche, ensuite les approches les plus connues pour la segmentation et le regroupement. Enfin, quelques systèmes « complets » de segmentation et regroupement en locuteurs sont présentés.

2.2. Définition du chant

Le **chant** est défini dans le dictionnaire le Petit Robert (1977) comme une « *émission des sons musicaux par la voix humaine (technique, art de la musique vocale)* ». Le dictionnaire des mots de la musique le définit comme une « *utilisation mélodique de la voix, associée ou non à des paroles, art de la voix mélodique* ». Ces deux définitions associent le chant à la musique vocale et considèrent que le chant ne peut être produit que par la voix. Le Trésor de la Langue Française définit aussi le chant comme une musique vocale : « *intonation particulière de même nature que celle de la parole, à la différence que dans le chant, la voix s'élève et s'infléchit bien davantage en modulant sur les différents degrés de l'échelle diatonique accessibles au registre du chanteur* », mais il le considère aussi, dans certains cas, comme une musique instrumentale produite par des instruments telle que le chant de la flûte et le chant des guitares.

Comme nous nous sommes intéressés à identifier les alternances entre chanteurs, nous nous basons sur les définitions qui considèrent que le chant est produit par la voix humaine (chanteurs).

Dans le domaine du traitement automatique du chant, plusieurs travaux ont été effectués tels que la détection de la voix chantée (Lachambre, et al., 2009), l'identification de chanteur (Shen, et al., 2006) ou encore la segmentation du chant en solo/chœur (Le Coz, et al., 2012). La plupart de ces travaux sont dans la continuité de ceux réalisés en reconnaissance des instruments dans la communauté du traitement de la musique ou de ceux effectués en détection de la parole et de la musique dans la communauté du traitement du signal. Le traitement du chant est un objectif commun pour les chercheurs en traitement de la parole et les chercheurs en traitement de la musique. Cela est dû aux caractéristiques du chant qui le rapprochent à la fois de celles de la parole par son mode de production : ils sont, tous les deux, produits par le même instrument (la voix humaine) (Castellengo, 2007), (Henrich Bernardoni, 2014) en suivant les mêmes lois physiques et physiologiques, et à celles de la musique puisqu'il est souvent accompagné par d'autres instruments. Cela peut conduire à considérer le chant comme un son intermédiaire entre de la parole et de la musique (instrumentale).

Malgré les points communs entre le chant et la parole, il existe des points de divergence que Nathalie Henrich mentionne dans ses recherches sur la voix chantée (Henrich Bernardoni, 2014a). En effet, une maîtrise consciente et même savante du contrôle de la pression sous-glottique (pression du souffle), de la hauteur de la voix et de sa dynamique est nécessaire pour générer de la voix chantée. Ce contrôle a une influence sur d'autres paramètres tels que le *jitter*, le vibrato et le trémolo.

2.3. Le traitement Automatique du Chant

Les travaux effectués sur le chant diffèrent à plusieurs niveaux : l'objectif, les paramètres et le contexte d'étude. Dans cette section, nous présentons trois systèmes de traitement du chant différents correspondant à trois objectifs différents : détection du chant, segmentation du chant en solo/chœur et structuration musicale. Nous compléterons par la description de quelques paramètres couramment utilisés pour le chant.

2.3.1. Détection de chant

La détection de chant consiste à localiser les segments correspondant à la présence de chant dans un morceau de musique. Plusieurs méthodes ont été proposées comme la méthode décrite dans (Berenzweig, et al., 2001) qui utilise des Modèles de Markov Cachés (Hidden Markov Models – HMM), qui permettent de prendre en compte la durée de chant. Les auteurs de cette méthode sont partis de l'idée que la parole et le chant possèdent des caractéristiques communes (structure formantique, transition entre phonèmes) et ils utilisent les probabilités *a posteriori* apprises sur un modèle de parole comme paramètre pour la détection de la voix chantée. D'autres approches basées sur des méthodes de décision Bayésiennes avec des Mélanges de Lois Gaussiennes (Gaussian Mixture Models – GMM) pour la modélisation de la distribution des paramètres, ont été développées : dans (Lukashevich, et al., 2007), un algorithme de détection automatique des régions de chant dans la musique populaire a été implémenté en utilisant les MFCC, avec les GMM, pour discriminer deux classes de données

(classe de musique instrumentale et classe de la voix chantée avec un fond musical). Un post-traitement supplémentaire a été proposé dans ces travaux pour améliorer les résultats de classification. Dans (Ezzaidi, et al., 2002), deux systèmes de discrimination de la parole, de la musique et du chant ont été proposés. Le premier utilise 3 GMM : un pour la parole, un pour la musique et un pour le chant. Chaque GMM est composé de 8 gaussiennes apprises sur de très courtes sessions. Le deuxième système est basé sur une comparaison des distances entre les dérivées des paramètres ΔMFCC à l'aide d'un seuil prédéfini.

Dans (Lachambre, et al., 2009a), une méthode de détection du chant basée sur la détection du vibrato a été proposée. Dans ce travail, une prise en compte du contexte monophonique (une seule source harmonique) et polyphonique (plusieurs sources harmoniques simultanées) a été faite afin d'améliorer le résultat du système de détection lorsqu'il s'agit d'un son polyphonique. De fait, un module de prétraitement est ajouté pour effectuer la séparation monophonie/polyphonie en utilisant comme paramètre un estimateur de fréquence fondamentale appelé YIN proposé par (De Cheveigné, et al., 2002). Cet estimateur génère en sortie une valeur interprétée comme un indice de confiance ; la moyenne et la variance à court terme de cet indice sont calculées et elles sont modélisées par la suite via la méthode des moments en utilisant des modèles de Weibull bivariés pour pouvoir effectuer la classification en monophonie/polyphonie.

Pour détecter du chant dans le cas d'un signal monophonique, le système de (Lachambre, et al., 2009a) utilise le vibrato, oscillation périodique de la fréquence fondamentale, qui a la particularité d'être toujours présente à une fréquence entre 4 et 8 Hz pour le chant. Dans le cas d'un son polyphonique, un autre traitement a été mis en œuvre qui consiste tout d'abord à effectuer une segmentation sinusoïdale, qui a été proposée par (Taniguchi, et al., 2005), basée sur un suivi temporel des fréquences positionnées dans l'espace temps-fréquence. De ces segments sinusoïdaux sont extraits pour chacun quatre paramètres : l'indice de début, l'indice de fin, le vecteur de suivi des fréquences et celui des amplitudes. Pour la recherche de ces segments, dans un premier temps, un calcul et un lissage du spectre sont réalisés. Dans un second temps, une conversion des fréquences en *cent* est effectuée et elle est suivie, par la suite, d'une détection des maxima du spectre. Dans un dernier temps, les distances entre les différents maxima du spectre sont calculées et les points, dont la distance est inférieure à un certain seuil, sont reliés entre eux.

Les relations temporelles entre les segments sinusoïdaux permettent de définir un segment temporel cohérent d'un point de vue harmonique. Le vibrato étendu, qui a été proposé dans (Lachambre, 2008), est calculé en déterminant, pour un segment temporel, la proportion de segments sinusoïdaux qui possèdent du vibrato. Les auteurs considèrent qu'il y a du chant si la valeur du vibrato étendu est suffisamment élevée.

2.3.2. Segmentation du chant en solo/chœur

Dans un contexte d'indexation d'un document musical et en présence de plusieurs chanteurs, la question suivante peut se poser : « est-on en présence d'un segment de chant

d'un chœur ou d'un solo ? ». Les travaux décrits dans (Le Coz, et al., 2012) proposent un système de distinction chœur-solo.

Ce système est composé de quatre étapes : localisation temporelle, localisation fréquentielle, suivi fréquentiel et classification solo/chœur. Les deux premières étapes de localisation temporelle et fréquentielle permettent de déterminer les zones d'intérêt dans le signal afin de ne pas traiter les zones bruitées ou de silence sur lesquels l'algorithme proposé ne pourrait donner une quelconque réponse.

La première étape de localisation temporelle consiste à localiser les longues phases stables dans le signal en utilisant une segmentation appelée segmentation *Forward-Backward*, qui a été développée par André-Obrecht (André-Obrecht, 1988) et qui permet de détecter les zones quasi-stationnaires du signal. L'application de cette segmentation sur la musique a montré un pouvoir d'isolation des quatre phases constituant une note (Le Coz, et al., 2010) : *Attack*, *Decay*, *Sustain*, *Release*. Son utilisation dans cette première étape permet de trouver les segments longs qui correspondent aux phases de *Sustain* ; ces segments sont particulièrement intéressants car, si le signal est produit par plusieurs chanteurs (chœur), il doit apparaître une divergence entre les harmoniques. Une sélection finale de ces segments est effectuée en ne considérant que ceux qui possèdent une valeur de la moyenne du critère de confiance de l'algorithme YIN (De Cheveigné, et al., 2002) supérieure à un certain seuil.

Après avoir extrait les zones temporelles d'intérêt dans lesquelles le phénomène de divergence entre harmoniques peut exister, une deuxième étape de localisation fréquentielle est effectuée pour limiter le nombre de bandes de fréquences qui seront traitées par la suite du système. La série des N bandes sur lesquelles l'analyse sera effectuée par la suite, est déterminé à partir d'une fonction d'estimation de la fréquence fondamentale corrigée $f_0(t)$. Cette fonction d'estimation consiste, tout d'abord, à calculer la fréquence fondamentale sur chaque trame en appliquant l'algorithme d'extraction de fréquence fondamentale YIN. Ensuite, un lissage médian est réalisé sur les estimations de fréquence fondamentale. Enfin, les sauts d'octaves sont corrigés en se basant sur la « fréquence » majoritaire dans le segment. Chaque bande de fréquence i de la trame t est centrée autour de la valeur $i \times f_0(t)$ pour l'approcher de la $i^{ème}$ harmonique et sa largeur est définie en fonction du rang de l'harmonique. Les deux segments conservés sont ceux les plus grands par seconde.

La troisième étape du suivi de fréquences qui permet de distinguer une divergence propre au phénomène de chœur, consiste à chercher les pics spectraux appartenant à deux trames consécutives qui correspondent à l'évolution d'un même phénomène et de le localiser par la suite dans les deux domaines fréquentiel et temporel. L'approche utilisée est celle des segments sinusoïdaux qui a été proposée par (Taniguchi, et al., 2005) et qui a servi à la détection du chant décrite dans la section précédente. Les pics qui sont conservés sont ceux qui se trouvent dans un voisinage proche de l'espace temps-fréquence et qui appartiennent à un même segment sinusoïdal. Une sélection des segments valides contenant des informations significatives, est effectuée en ne considérant que les segments suffisamment longs et qui ont une longueur supérieure à un certain seuil.

La dernière étape est la classification en solo/chœur et qui consiste à déterminer pour chaque segment s'il a été produit par un chanteur (solo) ou par plusieurs (chœur). Un taux de dédoublement des segments sinusoïdaux est calculé sur l'ensemble du segment d'analyse. Ce taux sera comparé à deux seuils, déterminés expérimentalement, pour décider si le segment d'analyse est considéré comme produit par un solo ou par un chœur.

2.3.3. Segmentation en structure musicale

En partant du fait que le chant est classé dans la catégorie Musique par la plupart des systèmes de classification Parole/Musique, nous avons étudié quelques systèmes de traitement de la musique. Parmi ces systèmes, nous nous sommes intéressés aux systèmes de structuration musicale puisque la tâche envisagée dans notre thèse peut s'apparenter à une recherche de structuration basée sur la présence de plusieurs chanteurs.

La segmentation à des fins de structuration musicale vise à extraire des informations pour définir des entités structurelles qui composent un morceau de musique, à savoir les frontières des segments, la forme musicale, et les étiquettes sémantiques comme le couplet, le refrain, etc. Ces informations extraites peuvent être utilisées pour créer des extraits représentatifs de chansons ou des résumés, afin de faciliter la navigation dans les grandes collections de musique ou comme prétraitement pour faciliter des applications de traitement plus avancé des enregistrements musicaux.

Il existe plusieurs méthodes d'analyse en structuration musicale, dont l'objectif est généralement de découper un enregistrement audio en des segments temporels correspondants aux parties musicales, puis de regrouper ces segments en des catégories musicalement significatives. Trois différentes approches peuvent être identifiées selon des critères de segmentation et de structuration : l'approche basée sur la répétition, l'approche basée sur la « Nouveauté » et l'approche basée sur l'homogénéité. D'autres dimensions sont à prendre en considération comme la mélodie, l'harmonicité, le rythme et le timbre.

Dans (Kaiser, et al., 2014) par exemple, le système de segmentation en structure musicale utilisé est basé sur la « Nouveauté » et l'homogénéité. Son architecture est présentée dans la Figure 2. 1, qui est une architecture commune à la plupart des systèmes d'analyse de structure musicale. Il est composé de cinq étapes : la première consiste à extraire les paramètres acoustiques qui sont généralement les chroma. Ensuite, le contexte tonal est déduit à partir d'une modélisation des séquences de trames des chroma locaux avec des « Multi-Probe Histogram » (MPH) qui permettent d'avoir une représentation statistique de la corrélation entre les séquences acoustiques. Le MPH est calculé pour chaque séquence de vecteurs-chromas et il conserve les informations sur les corrélations temporelles locales. Plus de détails sur le calcul des MPH sont décrits dans (Yu, et al., 2010). Ces histogrammes reflètent la structure tonale des parties locales de l'audio et permettent de modéliser l'évolution du contexte tonal et de capturer les lentes variations des motifs harmoniques reliés aux motifs musicaux. Cette description offre une bonne caractérisation de l'homogénéité des sections structurelles et permet aussi de discriminer les sections non liées entre elles. L'intégration des MPH dans une matrice de similarité renforce la visualisation de la structure. Ceci a été illustré

dans (Kaiser, et al., 2014) : en faisant l'extraction de la matrice de similarité à partir d'un exemple de chanson, des blocs de haute similarité, indiquant la présence des sections tonales homogènes à long terme et une représentation structurelle claire ont été obtenus.

En se basant sur le fait que les transitions entre les sections sont caractérisées par un changement significatif d'un contenu acoustique homogène à un autre contenu acoustique homogène, l'hypothèse de la méthode de segmentation temporelle suppose que les frontières dans un signal audio sont visualisées à partir de la matrice de similarité, qui est considérée comme un damier à deux dimensions.

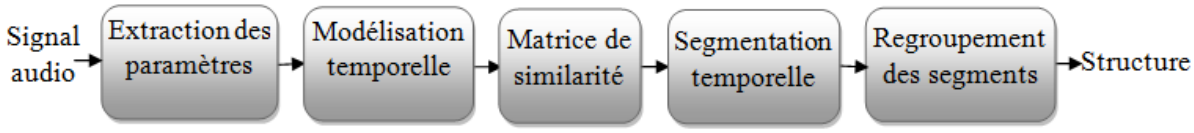

Figure 2. 1 - Architecture d'un système de segmentation en structure musicale.

Une méthode, basée sur la « KernelCorrelation », dont les détails sont décrits dans (Foote, 2000), est alors utilisée pour mieux modéliser la transition entre deux différents états de structure musicale et elle permet de calculer une courbe de « Nouveauté » qui sert pour la détection des frontières.

Une fois la segmentation temporelle réalisée, un regroupement des segments est effectué en se basant sur le fait que l'information dans les matrices de similarité est très redondante dans le temps. La méthode de regroupement utilisée effectue une réduction de la dimension de la matrice de similarité au moyen d'une Factorisation par Matrice Non-négative et permet par la suite de regrouper les segments en fonction de la structure musicale en appliquant un regroupement hiérarchique sur les segments projetés dans cette nouvelle base.

2.3.4. Paramétrisation

Quelque soit le système de segmentation d'un signal sonore, la phase de paramétrisation est essentielle et elle a pour but l'extraction d'informations pertinentes et discriminantes en fonction de la tâche envisagée.

- **MFCC**

Les coefficients cepstraux (MFCC – Mel Frequency Cepstral Coefficient) sont les paramètres les plus utilisés en traitement de la parole et aussi en traitement de la musique. Les détails de calcul de ces paramètres sont présentés dans la section suivante de segmentation et regroupement en locuteurs. Le chant étant considéré comme un son dérivant à la fois de la parole et de la musique, de nombreux travaux sur la voix chantée se fondent sur les MFCC. Dans (Lukashevich, et al., 2007), les auteurs utilisent un modèle de décision basé sur une modélisation de type GMM apprise sur des MFCC pour la détection de la voix chantée. Les MFCC ont été comparés avec la plupart des paramètres classiques utilisés en reconnaissance de la parole et ils ont été choisis comme étant les meilleures caractéristiques pour détecter la

voix chantée (Rocamora, et al., 2007). L'utilisation d'une classification par des SVM (Support Vector Machines) appris sur des MFCC donne aussi des bonnes performances pour une classification pour des segments d'une seconde. D'autres approches utilisent les MFCC pour améliorer la performance de leurs systèmes tels que la méthode de détection de la voix chantée en utilisant des paramètres de modulation de fréquence développée dans (Markaki, et al., 2008).

Quelques études en traitement du chant se sont servies des paramètres couramment utilisés en traitement du signal. Nous présentons quelques paramètres dans cette section.

- **ZCR**

Le ZCR est le taux de passage par zéro du signal. Le ZCR est un paramètre temporel fréquemment utilisé pour la segmentation parole/musique (Saunders, 1996) et (Scheirer, et al., 1997). La présence de la parole entraine des variations brusques du ZCR qui sont significatives de l'alternance voisée/non voisée. Par contre, pour la musique le ZCR varie faiblement. Il est notamment élevé pour un signal bruité. Ce paramètres a été aussi utilisé dans des travaux en traitement du chant tels que ceux décrits dans (Rocamora, et al., 2007) pour la détection de la voix chantée, dans (Zhang, 2002) pour un système d'identification automatique de chanteurs où le ZCR est utilisé avec d'autres paramètres pour localiser les points de début de la voix chantée dans une chanson et dans les travaux de (Ramona, et al., 2008) pour la détection vocale en musique avec les SVM. Son mode de calcul est le suivant :

Une trame acoustique est une suite d'échantillons représentant de 20 à 40 ms du signal durant laquelle le signal est supposé quasi stationnaire. Le ZCR d'une trame est calculé à partir du nombre de changements de signe du signal sonore :

$$ZCR(i) = \frac{1}{2N}\left(\sum_{n=1}^{N}\left|sign\left(x_n(i)\right) - sign\left(x_{n-1}(i)\right)\right|\right) \qquad (2.1)$$

avec $x_n(i)$ le n^{ième} échantillon de la trame i et $(N+1)$ le nombre total des échantillons dans la trame i.

Une étude, menée par (Lu, et al., 2001), a montré que la variation du ZCR est plus significative que sa valeur exacte. Il en a déduit la mesure « haute proportion du ZCR » (HZCRR). Cette mesure correspond au nombre où le ZCR est supérieur à 1,5 fois la valeur moyenne du ZCR, par unité de temps.

$$HZCRR = \frac{1}{2I}\left(\sum_{i=1}^{I}\left(sign(ZCR(i) - 1.5 * avZCR) + 1\right)\right) \qquad (2.2)$$

avec I le nombre total de trames et

$$avZCR = \frac{1}{I}\sum_{i=1}^{I} ZCR(i) \qquad (2.3)$$

- **Energie**

L'énergie est un paramètre temporel fréquemment utilisé en traitement du signal. Ce paramètre a prouvé son pouvoir de segmentation en parole / bruit et de détection de silence. L'énergie à très court terme est plutôt stable pour la musique alors qu'elle est très variable pour la parole. L'énergie globale d'un signal échantillonné de longueur finie $\left(x_n(i)\right)_{n=1,\dots,N}$ est définie par :

$$E(i) = \sum_{n=1}^{N} x_n^2(i) \tag{2.4}$$

Afin de respecter l'échelle perceptive, l'énergie est généralement exprimée en décibels :

$$E_{db}(i) = 10 \times \log_{10}\left(\sum_{n=1}^{N} x_n^2(i)\right) \tag{2.5}$$

Pour un signal échantillonné à support infini, l'énergie à court terme est calculée en prenant des portions de signal relatives à une fenêtre glissante. Cette fenêtre correspond en général à une trame acoustique de l'ordre de 10 ms.

Une des causes de la variabilité de ce paramètre est due à des conditions d'enregistrements différentes. En effet, une simple variation de la distance entre la source et le microphone suffit pour être un élément de perturbation de l'énergie. Afin d'éliminer ce problème de variabilité, une normalisation de l'énergie par rapport au maximum observé sur le signal global peut être effectuée.

L'énergie a été aussi utilisée dans des travaux sur le chant. Dans (Zhang, 2003), elle a été accompagnée du ZCR et d'autres paramètres pour trouver les points de début des segments du chant qui vont être utilisés par la suite pour l'identification automatique des chanteurs. Une méthode d'identification de la voix chantée a été développée dans (Mesaros, et al., 2006) en utilisant les coefficients d'énergie comme paramètres. Les auteurs de cette approche utilisent deux représentations de l'énergie du signal vocal et évaluent leur efficacité pour l'identification de la voix chantée. La première représentation est composée des énergies dans l'échelle du Mel des 14 bandes de fréquences couvrant tout le spectre fréquentiel du signal. La deuxième représentation est obtenue par une décomposition en ondelettes du signal.

- **Centroïde spectral**

Le centroïde spectral est un paramètre fréquentiel qui correspond au centre de gravité du spectre d'une DSP (Densité Spectrale de Puissance) et il est défini par :

$$C(i) = \frac{\sum_{n=1}^{N} w_n S_i(w_n)}{\sum_{n=1}^{N} S_i(w_n)} \tag{2.6}$$

où $S_i(w_n)$ est l'amplitude de la n^{ième} composante fréquentielle w_n de la trame i et N est le nombre de composantes fréquentielles dans la trame i.

Le centroïde spectral est plus élevé pour la musique car l'intensité des harmoniques des sons reste prépondérante sur une zone fréquentielle plus importante pour la musique que celle pour la parole (atténuation de l'excitation vocale). De plus, une variation importante de ce paramètre caractérise l'alternance voisée/non-voisée. Il a été aussi utilisé avec le ZCR et l'énergie dans (Rocamora, et al., 2007) pour la détection de la voix chantée.

- **Flux spectral**

Le flux spectral est un paramètre fréquentiel qui mesure les variations à court terme du spectre et sa dynamique. Il correspond à la distance euclidienne entre deux transformées de Fourier calculées sur des trames successives.

$$FS = \sum_{n=1}^{N} \left(\frac{S_i(w_n)}{\|S_i\|} - \frac{S_{i-1}(w_n)}{\|S_{i-1}\|} \right)^2 \tag{2.7}$$

où $S_i(w_n)$ correspond à la composante spectrale de la trame i à la fréquence w_n.

Pour la musique, la valeur du flux spectral est élevée et les variations sont faibles. Par contre pour la parole, les variations sont importantes et le niveau peut être faible. L'accompagnement de ce paramètre avec le ZCR, l'énergie et le coefficient harmonique, qui est détaillé dans le paragraphe suivant, a aidé les auteurs de (Zhang, 2003) à avoir des résultats prometteurs pour identifier automatiquement les chanteurs.

- **Coefficient harmonique**

Le coefficient harmonique proposé dans (Cho, et al., 1998) est un paramètre mixte, c'est-à-dire issu à la fois d'une analyse fréquentielle et temporelle. Il permet de mesurer le poids de la plus importante série harmonique dans une décomposition en somme de séries. La présence d'une ou plusieurs séries harmoniques est indiquée par une valeur élevée du coefficient harmonique H. Il est défini par :

$$H = \max_{\tau} R(\tau) \tag{2.8}$$

où $R(\tau)$ est une combinaison de deux fonctions d'autocorrélation : une fonction d'autocorrélation temporelle $R^T(\tau)$ et une fréquentielle $R^F(f_\tau)$ qui sont calculées de la manière suivante :

$$R^T(\tau) = \frac{\sum_{\mathrm{k}=0}^{\mathrm{K}-\tau-1} [\tilde{\mathrm{s}}(\mathrm{k}) \cdot \tilde{\mathrm{s}}(\mathrm{k}+\tau)]}{\sqrt{\sum_{\mathrm{k}=0}^{\mathrm{K}-\tau-1} \tilde{\mathrm{s}}^2(\mathrm{k}) \cdot \sum_{\mathrm{k}=0}^{\mathrm{K}-\tau-1} \tilde{\mathrm{s}}^2(\mathrm{k}+\tau)}} \tag{2.9}$$

$$R^F(f_\tau) = \frac{\sum_{\mathrm{f}=0}^{\mathrm{K}-f_\tau-1}\left[\tilde{\mathrm{S}}(\mathrm{f}) \cdot \tilde{\mathrm{S}}(\mathrm{f}+f_\tau)\right]}{\sqrt{\sum_{\mathrm{f}=0}^{\mathrm{K}-f_\tau-1} \tilde{\mathrm{S}}^2(\mathrm{f}) \cdot \sum_{\mathrm{f}=0}^{\mathrm{K}-f_\tau-1} \tilde{\mathrm{S}}^2(\mathrm{f}+f_\tau)}} \qquad (2.10)$$

où s est le signal à analyser, S est le module de sa transformée Fourier, $\tilde{s}$ est le signal s centré en zéro, $\tilde{S}$ est le module centré de sa transformée de Fourier et $f_\tau = {}^{2\pi K}/_{\tau}$.

D'autres paramètres musicaux ont été aussi utilisés pour le traitement du chant et l'analyse de la voix chantée. Parmi ces paramètres, il y a le timbre, le vibrato, le tempo et les chroma que nous présentons dans la suite de cette section.

- **Timbre**

Le timbre permet de faire la différence entre deux sons de même hauteur, puissance et durée (Wold, et al., 1999) et (Zhang, et al., 1998). Le timbre est une qualité spécifique des sons produits par un instrument : il traduit à la fois la répartition des harmoniques à un instant donné et il correspond à l'évolution temporelle du spectre dans la réalisation d'un son. L'ANSI (ANSI, 1960) (American National Standards Institute) définit le timbre comme la « *caractéristique sonore, qui permet à un auditeur de juger que deux sons présentés de la même façon, de même hauteur, durée et intensité sont différents* ». Donc, il est jugé comme une caractéristique spécifique d'un son vocal produit indépendamment de sa hauteur, sa durée et son intensité.

Marozeau, dans sa thèse (Marozeau, 2004), mentionne que la notion de timbre n'est pas encore bien comprise et que le timbre ne peut pas être décrit par une grandeur physique unique parce qu'il dépend de plusieurs facteurs perceptifs et qu'il peut évoluer en fonction du contexte. L'ANSI précise que « *Le timbre dépend premièrement du spectre fréquentiel du stimulus, mais également de la forme d'onde, de la pression acoustique, de la disposition des fréquences à l'intérieur du spectre et des caractéristiques temporelles du stimulus* ».

Dans la littérature, la reconnaissance du timbre était la problématique principale des recherches menées sur la tâche d'identification des instruments. La reconnaissance des instruments était, au départ, étudiée à partir d'une seule note produite par des instruments solos, puis à partir de toute une phrase musicale.

Des études réalisées dans (Pols, et al., 1969) ont montré que la prédiction du timbre de sons complexes par une grandeur physique est possible en trouvant des bonnes corrélations entre les dimensions perceptives et physiques. Dans cette étude, les auteurs ont effectué une représentation des caractéristiques fréquentielles et temporelles des stimuli dans un espace en calculant pour chaque stimulus l'énergie par bandes de fréquences. Une analyse par composante principale est réalisée par la suite pour réduire la taille des données obtenues afin de trouver les corrélations entre les dimensions physiques et perceptives.

- **Vibrato**

Dans (Seashore, 1938), le vibrato a été défini comme « une oscillation de la fréquence fondamentale, habituellement accompagnée de variations synchrones de la puissance et du timbre, dont l'étendue et la fréquence sont telles qu'elles ajoutent au son une flexibilité, une tendresse et une richesse plaisante ». Donc, le vibrato est l'oscillation périodique de la fréquence fondamentale. Cette oscillation peut être produite par un chanteur ou par un instrument et elle peut se faire soit sur la valeur même de la fréquence, soit sur l'intensité. Le vibrato est appelé « trémolo » quand l'oscillation se fait sur l'intensité du son (Regnier, et al., 2009).

Pour les instruments, le vibrato est un effet musical ajouté volontairement par le musicien qui choisit la fréquence d'oscillation. Son mode de production dépend de l'instrument utilisé (Timmers, et al., 2000). Pour le violon par exemple, le vibrato est produit par le mouvement régulier de l'oscillation périodique de la main gauche sur le manche de l'instrument, ce qui implique qu'il y a un changement de la valeur de la fréquence fondamentale, par contre l'intensité ne change que lorsque la pression de l'archet sur les cordes varie.

Dans le cas du chant, le vibrato est produit spontanément par la voix humaine et les oscillations se font sur la valeur de la fréquence fondamentale et correspondent toujours à un rythme compris entre 4 et 8 Hz (Seashore, 1938), (Sundberg, 1994), (Meron, et al., 2000), (Rossignol, et al., 1999). Les chanteurs professionnels peuvent l'amplifier ou l'atténuer mais ils ne peuvent pas l'enlever lorsqu'ils chantent. Pour un chanteur donné, le rythme des oscillations est souvent constant et il peut être utilisé pour identifier un chanteur. Néanmoins, les oscillations fréquentielles possèdent une étendue très variable qui peut atteindre le demi-ton, même pour un chanteur.

Dans le système de détection de la voix chantée de Lachambre (Lachambre, et al., 2009a) qui a été décrit dans la partie 2.3.1, le vibrato a été utilisé pour localiser les segments de chant et le « vibrato étendu », synonyme de présence d'oscillations sur l'ensemble des harmoniques, a permis de rendre plus robuste cette détection.

- **Tempo**

Le tempo est l'une des caractéristiques principales d'un morceau de musique. Il permet d'avoir une information sur la vitesse d'exécution d'un morceau et aussi de localiser les pulsations et donc les temps : l'unité de la structure musicale.

Les pulsations sont les instants de coupure possibles qui correspondent aux changements dans la structure d'un morceau. Elles correspondent à des accentuations données à la musique de manière cyclique, garantissant que toutes les unités possèdent la même période. Le tempo est la fréquence correspondante à cette période.

Le tempo considéré dans (Peeters, 2007) comme étant le **tactus** sachant que ce dernier est défini comme « *la période perceptuellement la plus dominante, […] est la fréquence à laquelle la majorité des gens taperaient du pied ou des mains en phase avec la musique* ». La perception du tempo diffère d'une personne à l'autre et dépend de plusieurs facteurs : l'âge, les connaissances musicales et le moment de la journée. Néanmoins, pour le même morceau de musique, la valeur du tempo varie d'un rapport de deux, un-demi, trois ou un-tiers si l'un de ces facteurs change. Ce paramètre est généralement utilisé pour les tâches d'indexation musicale et d'analyse automatique de la musique.

Le tempo est utilisé pour la détection du rythme par l'analyse de signaux audio polyphoniques en se basant généralement sur une approche qui effectue une analyse directe du signal pour extraire le tactus. Cette approche est la plus courante et sert aussi à extraire d'autres paramètres : le **tatum** et la **mesure** et elle est constituée de quatre étapes. La première est la bonne représentation du signal et les représentations généralement utilisées pour cette étape sont l'énergie par bandes de fréquences (Scheirer, 1997a), la représentation temps-fréquence (Klapuri, 2006) et la Transformée de Fourier Discrète (Peeters, 2005). La deuxième étape consiste à comparer les vecteurs d'observation à différents instants par une fonction de différence (Scheirer, 1997a), de similarité (Foote, et al., 2001) ou d'autocorrélation (Peeters, 2005). La troisième étape effectue une analyse fréquentielle de la fonction précédente. La dernière étape permet de déduire la pulsation et donc le tempo.

- **Chroma**

Les chromas sont des paramètres couramment utilisés en traitement de la musique pour des buts de classification par exemple : classification des artistes par le genre ou par le style ou classification des instruments (Daniel, et al., 2007). Un vecteur chroma est composé de 12 éléments et chaque élément représente le niveau sonore de chaque note (l'intensité d'un demi-ton) de la gamme dans un accord donné. L'importance de chaque demi-ton est modélisée par les « Key profile », appelées aussi les « profils de clé ».

Il s'agit d'une représentation fréquentielle sous une forme circulaire qui a été proposée par Shepard (Shepard, 1964). Le $n^{ième}$ coefficient du vecteur chroma représente l'énergie cumulée dans les bandes de fréquences correspondantes à la $n^{ième}$ note de la gamme, sur toutes les octaves possibles. Dans la littérature, il existe plusieurs types de chroma qui diffèrent dans leur mode de calcul. La première étape de calcul est commune pour tous les types et elle consiste à faire une analyse temps-fréquence du signal par Transformée de Fourier à Court Terme (TFCT). Ensuite, pour le calcul des vecteurs, différentes procédures ont été proposées. Parmi ces procédures, nous trouvons :

- Les « *Pitch Class Profiles* » (Fujishima, 1999) sont calculés en sommant les points de TFCT en fonction du chroma correspondant à leur fréquence associé.

- Les « *CQ-profiles* » (Purwins, 2005) effectuent une décomposition du signal par un banc de filtres centrés sur les demi-tons de la gamme chromatique, à Q constant. Q

représente le rapport entre la fréquence centrale de chaque filtre et la largeur de sa bande passante. Les coefficients *CQ-profiles* sont obtenus en sommant les composantes du signal correspondant à chaque chroma.

Les « chroma vectors » sont utilisés dans plusieurs travaux sur l'identification de la tonalité. Dans (Izmirli, 2005) et (Gomez, 2006), par exemple, une modélisation de la répartition des « chroma vectors » pour chacune de tonalités est effectuée. Ensuite, une comparaison de la moyenne des « chroma vectors » à chacun des modèles est réalisée pour estimer la tonalité d'un extrait musical.

2.3.5. Conclusion

Il existe de nombreux travaux effectués sur le chant, ainsi qu'une grande variété de paramétrisations possibles pour détecter le chant, extraire ses caractéristiques telles que la tonalité, et identifier son producteur (chanteur). Mais à notre connaissance, il n'existe pas d'études sur la tâche envisagée de segmentation et regroupement en tours de chant ou encore appelée segmentation et regroupement en chanteurs. En parole, de nombreux travaux sur la segmentation et regroupement en locuteurs ont été réalisés. En exploitant le fait que le chant et la parole partagent certaines caractéristiques, nous avons procédé par analogie pour segmenter et regrouper en tours de chant. D'où la nécessité d'effectuer une étude des systèmes de segmentation et regroupement en locuteurs et de dresser un état de l'art présenté dans la partie suivante.

2.4. Segmentation et regroupement en locuteurs

La Segmentation et Regroupement en Locuteurs (SRL), connu aussi sous le nom de « *speaker diarization* » en anglais, consiste tout d'abord à découper un enregistrement audio en tour de parole en le segmentant en zones acoustiquement homogènes par locuteur. Ensuite, les segments qui ont été prononcés par le même locuteur sont regroupés pour former une même classe afin de répondre à la question « **qui parle et quand ?** ». La SRL est très utile dans de nombreux types d'applications :

- Indexation des bases de données audio,
- Transcription automatique de la parole : la fusion des segments du même locuteur dans un document audio permet l'augmentation des données utilisées pour l'adaptation non supervisée des modèles aux locuteurs, améliorant ainsi les performances de la reconnaissance de la parole,
- Suivi de locuteur : dans les systèmes de suivi de locuteur, la vérification du locuteur cible se fait sur une dizaine de millisecondes. La SRL permet de réaliser cette vérification sur une quantité de données plus importante, rendant ainsi la décision plus robuste.

Plusieurs systèmes de segmentation et regroupement en locuteurs ont été développés (Anguera Miro, 2006), (El-Khoury, et al., 2009), (Meignier, et al., 2009), (Bost, et al., 2014), (Bost, et al., 2015). Dans la littérature, la plupart des travaux réalisés en SRL sont composés de trois blocs indispensables :

- Paramétrisation,
- Segmentation en tours de parole,
- Regroupement en locuteurs.

Dans cette partie, nous commençons par présenter, dans une première section, les paramètres acoustiques les plus utilisés en SRL. Dans une deuxième section, nous nous intéressons à l'étape de segmentation en tours de parole et nous présentons les types de systèmes de segmentation existant ainsi que les approches sur lesquelles ils se basent. Les principaux systèmes de regroupement en locuteurs ainsi que les différentes méthodes implémentés pour cette tâche, font l'objet d'une troisième section. Une quatrième section est consacrée à la présentation de quelques systèmes complets de segmentation et regroupement en locuteurs.

2.4.1. Paramétrisation

La phase de paramétrisation a pour but l'extraction des informations pertinentes des locuteurs à partir du signal audio enregistré durant leur conversation. Le choix des paramètres a une grande influence sur le processus de segmentation car ils représentent les informations qui permettent de trouver les spécificités qui caractérisent la voix de chaque locuteur et ainsi de le distinguer d'autres locuteurs.

Les paramètres les plus utilisés en traitement de la parole et plus particulièrement pour les techniques de traitement basées sur les locuteurs, sont les paramètres de type cepstral car ils permettent de rendre compte de la perception auditive grâce à l'échelle du Mel. Parmi ces paramètres, nous trouvons les coefficients cepstraux (Mel Frequency Cepstral Coefficient – MFCC) et les bancs de filtres (FBANK), qui sont très répandus dans les systèmes SRL. Il y a aussi les coefficients de prédiction linéaire perceptive (Perceptual Linear Prediction – PLP) et les RASTA-PLP (RelAtiveSpecTrAl – PLP) qui sont aussi utilisés pour l'analyse de la parole.

D'autres paramètres classiques du traitement du signal sont aussi utilisés dans les systèmes de segmentation et regroupement en locuteurs pour détecter l'activité vocale durant la phase de prétraitement. Parmi ces paramètres utilisés, nous trouvons l'énergie et le ZCR qui ont été détaillés dans la partie 2.3.4. De plus, la modulation d'énergie à 4 Hz qui est un paramètre mixte, s'est avérée utile pour la détection de la parole. En effet, la modulation d'énergie atteint son pic lorsqu'il s'agit des changements de syllabes, ce qui en fait un signal périodique dont la fréquence est autour de 4 Hz (4 syllabes par seconde) (Houtgast, et al., 1985). Cette propriété a été utilisée pour séparer la parole de la musique, pour distinguer la parole propre de la parole bruitée, ou la parole mono-locuteur de zones d'interactions où plusieurs locuteurs parlent simultanément.

Dans les parties suivantes, nous détaillons les paramètres acoustiques les plus utilisés dans les systèmes de traitement automatique de la parole : MFCC, FBANK, PLP et RASTA-PLP.

2.4.1.1. MFCC et FBANK

Les MFCC sont des paramètres très connus et utilisés non seulement en traitement de la parole, mais aussi dans d'autres domaines de l'analyse de l'audio telles l'analyse de la musique (Lukashevich, et al., 2007), (Rocamora, et al., 2007), (Markaki, et al., 2008). Ces paramètres sont bien adaptés au signal de la parole. Etant donné que la voix est considérée comme le produit de convolution entre les cordes vocales (la source) et le conduit vocal (le filtre), les MFCC, qui permettent de transformer un produit de convolution en une somme, séparent la source du conduit.

Le processus d'extraction des coefficients cepstraux est fondé sur la transformation de l'amplitude spectrale grâce à un banc de filtres (Calliope, 1989). Le banc de filtres est caractérisé par des filtres triangulaires, répartis de manière linéaire selon l'échelle Mel, défini dans l'équation (2.11), pour caractériser au mieux la perception humaine des sons. Ils sont linéairement espacés avec la même largeur de bande jusqu'à 1000 Hz. Ensuite, ils sont espacés d'une façon logarithmique. La transformation de l'amplitude spectrale consiste à intégrer les énergies spectrales par des fonctions de pondération d'un ensemble de bandes limitées par les filtres triangulaires.

$$Mel(f) = 2595\ log_{10}\left(1 + \frac{f}{700}\right) \tag{2.11}$$

Le logarithme des énergies obtenu après filtrage est finalement calculé pour avoir les coefficients FBANK. La Figure 2. 2 présente les différentes étapes de calcul des coefficients FBANK.

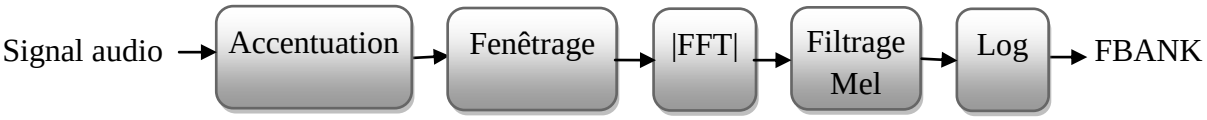

Figure 2. 2 - Les étapes de calcul des coefficients FBANK.

où FFT est la Transformée de Fourier Rapide qui se calcule sur des fenêtres glissantes.

Une accentuation des composantes fréquentielles aigues est effectuée car les composantes aigues sont plus faibles que les graves. Un filtrage passe-haut est réalisé avec la fonction de transfert suivante :

$$H(z) = 1 - 0.97 * z^{-1} \tag{2.12}$$

Afin d'éviter la formation des artefacts liés aux effets de bord, un fenêtrage de *Hamming* est réalisé sur chaque trame avec un recouvrement sur la moitié :

$$W_{Hamming}\ (n) = \begin{Bmatrix} 0.54 - 0.46 * \cos\left(\frac{2\pi n}{N}\right) \\ 0 \end{Bmatrix} \begin{vmatrix} pour\ 0 \leq n \leq N-1 \\ ailleurs \end{vmatrix} \quad (2.13)$$

avec N la taille de la fenêtre.

Afin d'obtenir les coefficients cepstraux (MFCC), une projection sur la base des cosinus doit être effectuée.

2.4.1.2. PLP

Les paramètres PLP (Perceptual Linear Prediction) sont, comme les MFCC, couramment adoptés pour l'analyse acoustique. La méthode de calcul de ces paramètres est détaillée dans la Figure 2. 3. Tout d'abord, un fenêtrage de *Hamming* défini dans l'équation (2.13) est effectué. Ensuite, l'amplitude spectrale du signal est calculée et suivie d'un filtrage selon l'échelle Bark. L'échelle Bark, comme l'échelle Mel, reproduit approximativement la sensibilité du système auditif humain. Puis, une pré-accentuation des composantes fréquentielles est appliquée. Ensuite, une accentuation du volume d'intensité sonore est appliquée par élévation à la puissance de 0,33. Un traitement par prédiction linéaire est réalisé sur le spectre obtenu. Enfin, les coefficients cepstraux sont obtenus par récurrence à partir des coefficients de prédiction, ce qui est l'équivalent du log du spectre suivi d'un transformé de Fourier inverse.

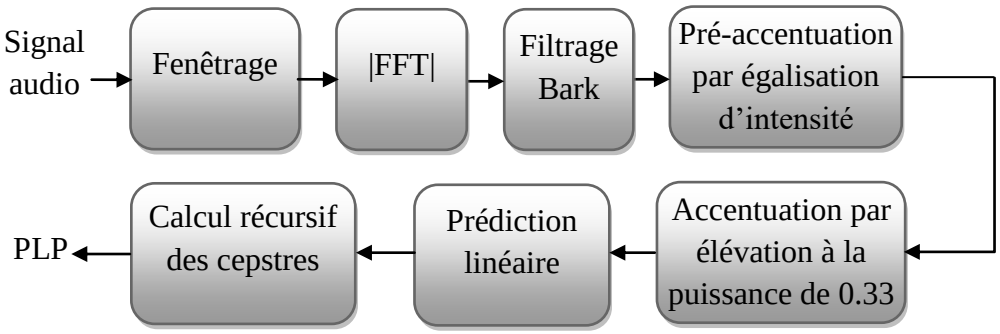

Figure 2. 3 - Les étapes de calcul des coefficients PLP.

2.4.1.3. RASTA-PLP

Les PLP ne sont pas trop robustes aux variations linéaires de la fonction de transfert et aux variations spectrales à long terme. Afin de remédier à ce problème, des techniques de normalisation spectrale ont été proposées. Parmi ces techniques proposées, une méthode appelée « RASTA-PLP » (RelAtiveSpecTrAl-PLP) a permis de réduire l'effet du bruit de convolution résultant des variations de la fonction de transfert.

La technique RASTA-PLP consiste à appliquer un filtrage temporel dans le domaine log-spectral avant la modélisation autorégressive (Hermansky, et al., 1991), (Boite, et al., 2000).

Dans le cas général, le processus de cette méthode peut être résumé par une équivalence avec un filtrage passe-bande de chaque bande de fréquence en utilisant un filtre IIR dont la fonction de transfert est définie par :

$$H(z) = 0.1 \times \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{z^{-4} \times (1 - 0.98z^{-1})} \tag{2.15}$$

Le changement spectral le plus rapide conservé est celui qui correspond à une haute fréquence de coupure de ce filtre, tandis que le changement spectral correspondant à une fréquence de coupure faible est ignoré.

2.4.2. Segmentation en tours de parole

Comme nous avons précisé dans le deuxième chapitre, les tours de parole n'existent que lorsqu'au moins deux locuteurs parlent à tour de rôle. La segmentation en tours de parole consiste à détecter dans un enregistrement audio les frontières de chaque tour de parole même en absence de silence entre deux locuteurs consécutifs. Cette détection se réalise en segmentant en locuteurs, c'est-à-dire en localisant les points de rupture entre locuteurs. La segmentation en locuteurs vise à diviser le flux audio en des segments acoustiquement homogènes par locuteur. La Détection de Changement de Locuteur est la méthode la plus utilisée pour la segmentation en locuteurs ou encore en tours de parole.

Les systèmes de détection de changement de locuteur existant dans la littérature peuvent être regroupés en deux types. Le premier concerne les systèmes qui effectuent une seule passe de traitement de l'enregistrement audio. Le second concerne les systèmes qui utilisent des algorithmes à plusieurs passes : dans la première étape, plusieurs points de changement sont proposés avec un taux de fausse alarme potentiellement élevé. Ensuite, pendant la deuxième étape, ces points sont réévalués et certains sont rejetés afin d'avoir une sortie optimale du système de segmentation en locuteurs.

Les méthodes présentées dans cette partie peuvent être classées en deux catégories : les approches basées sur une métrique comme la version symétrique de la distance Kullback-Leibler (KL2) et les approches basées sur un test d'hypothèses comme le Rapport de Vraisemblance Généralisé (Generalized Likelihood Ratio – GLR) et le Critère d'Information Bayésien (Bayesian Information Criterion – BIC).

2.4.2.1. La divergence de Kullback-Leibler symétrique

La distance de Kullback-Leibler (Kullback, et al., 1951) estime la dissimilarité entre deux distributions aléatoires. Pour deux distributions de probabilité continues p_1 et p_2 , la divergence de Kullback-Leibler de p_2 par rapport à p_1 est définie par la mesure suivante :

$$KL(p_1 \| p_2) = \int p_1(x) \log \frac{p_1(x)}{p_2(x)} dx \tag{2.16}$$

Cette distance est asymétrique par rapport aux deux variables p_1 et p_2, ce qui a conduit à proposer une version symétrique de cette divergence, appelée $KL2$:

$$KL2 = KL(p_1 \| p_2) + KL(p_2 \| p_1) \tag{2.17}$$

Lorsqu'il s'agit de deux distributions gaussiennes $N(\mu_1, \sigma_1^2)$ et $N(\mu_2, \sigma_2^2)$, l'expression de la divergence symétrique de Kullback-Leibler est de la forme suivante :

$$KL2 = \frac{1}{2}\left[\frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2} + (\mu_1 - \mu_2)^2\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)\right] \tag{2.18}$$

où μ_1, μ_2 représentent les moyennes de p_1 et p_2, σ_1^2 et σ_2^2 correspondent aux variances de p_1 et p_2, respectivement.

La méthode décrite dans (Siegler, et al., 1997) consiste à construire une courbe de distance avec les valeurs de $KL2$ calculées à partir de tous les points de l'enregistrement en considérant à chaque fois les deux fenêtres adjacentes. Le calcul de $KL2$ est facilement effectué après l'estimation de la moyenne et la variance pour chaque fenêtre. La détection des points de changement de locuteurs est réalisée en considérant les points correspondant aux maxima locaux de la courbe tracée.

2.4.2.2. Rapport de Vraisemblance Généralisé

Le Rapport de Vraisemblance Généralisé, proposé en parole par (Gish, et al., 1991), effectue une vérification des hypothèses en utilisant les segments acoustiques dans deux fenêtres consécutives glissantes et possiblement avec chevauchement. Pour chaque point de changement potentiel, il y a deux hypothèses possibles :

- la première H_0 suppose que, de part et d'autre de ce point, le signal obéit au même modèle probabiliste, noté $M_0(H_0)$,

- la deuxième H_1 suppose qu'il y a un changement de modèle et que deux modèles différents M_1 et M_2 sont nécessaires.

Dans la pratique, les modèles sont estimés sur trois fenêtres d'analyse et le critère GLR est utilisé pour déterminer si le signal est « mieux » représenté par deux modèles distincts ou par un modèle unique ; un seuil est déterminé de manière empirique ou adapté dynamiquement. Il s'en suit que si le signal analysé correspond à une séquence de N vecteurs d'observation (vecteurs acoustiques ou trames) de dimension d, noté $X_0(x_1, x_2, \dots, x_N)$; un point de changement potentiel positionné après la trame t induit deux sous suites consécutives $X_1(x_1, x_2, \dots, x_t)$ et $X_2(x_{t+1}, x_2, \dots, x_N)$. Le rapport de vraisemblance généralisé entre les deux hypothèses H_0 et H_1 est donné par :

$$GLR = \frac{P(H_0)}{P(H_1)} \tag{2.19}$$

Ce rapport est comparé à un seuil δ. Si le GLR est inférieur à δ, alors l'hypothèse H_1 est vérifiée, indiquant l'existence de deux modèles différents et donc un point de changement est détecté à l'instant t. En appliquant le log à l'expression du GLR donne :

$$R(t) = -log(GLR) \tag{2.20}$$

En supposant que X_0, X_1 et X_2 suivent des lois gaussiennes données respectivement par $M_0(\mu_0, \Sigma_0)$, $M_1(\mu_1, \Sigma_1)$ et $M_2(\mu_2, \Sigma_2)$, l'expression de $R(t)$ devient :

$$R(t) = \frac{1}{2}(N \log(|\widehat{\Sigma_0}|) - t \log(|\widehat{\Sigma_1}|) - (N-t) \log(|\widehat{\Sigma_2}|) \tag{2.21}$$

où $|\widehat{\Sigma_i}|$ est le déterminant de la matrice de covariance Σ_i, estimée sur la suite X_i avec $i = 0,1,2$. Le point de changement de locuteur est estimé à l'instant $\hat{t}$:

$$\hat{t} = \arg\max_t R(t) \tag{2.22}$$

Un point de changement de locuteur est détecté lorsque la valeur $\hat{t}$ dépasse le seuil $T = -\log \delta$, qui est déterminé sur un ensemble de développement.

2.4.2.3. Critère d'Information Bayésien

Le Critère d'Information Bayésien (BIC), comme le GLR, effectue une vérification des hypothèses et il considère les mêmes hypothèses H_0 et H_1 décrites dans la section 2.4.2.2. L'expression du BIC d'un modèle paramétrique M estimé à partir d'une séquence d'observation X est donnée par :

$$BIC(M) = \log L(X, M) - \frac{\lambda}{2} d \log N \tag{2.23}$$

où d est la dimension de vecteurs d'observation et N correspond au nombre de vecteurs d'observation du modèle. Le premier terme du BIC représente le log de la fonction de vraisemblance $L(X, M)$du modèle par rapport aux données. Le deuxième correspond à un terme de complexité ; λ est un coefficient de pénalité dont la valeur théorique et standard fixée par Rissanen (Rissanen, 1989), est égale à 1.

Une différence des expressions du BIC des deux modèles correspondant aux hypothèses H_0 et H_1 est effectuée pour avoir le critère de comparaison ΔBIC. En considérant des distributions gaussiennes, le ΔBIC à l'instant t est donné par :

$$\Delta BIC(t) = R(t) - \lambda P \tag{2.24}$$

où le log du rapport de vraisemblance $R(t)$ est défini par l'équation (2.21), et P est un terme de complexité proportionnel à la différence des nombres de paramètres estimés pour chaque hypothèse et vaut, dans le cas de matrices de covariance pleines :

$$P = \frac{1}{2}\left(d + \frac{1}{2}d(d+1)\right)\log N \qquad (2.25)$$

Dans le cas de matrices de covariance diagonales, l'expression du terme P devient :

$$P = \frac{1}{2}\, d\, \log N \qquad (2.26)$$

Le facteur de pénalité λ est appris de telle sorte que le critère ΔBIC soit positif dès lors que l'hypothèse H_1 est vérifiée, indiquant l'existence de deux modèles différents. Sinon, l'hypothèse H_0 est validée ainsi que l'existence d'un seul modèle pour la fenêtre X_0. Un changement est détecté dès lors que :

$$\{\max_t \Delta BIC(t) \geq 0\} \qquad (2.27)$$

La valeur estimée du point de changement est exprimée par :

$$\hat{t} = arg\max_t \Delta BIC(t) \qquad (2.28)$$

Dans (Chen, et al., 1998), les auteurs ont utilisé ce critère pour segmenter les données de la campagne d'évaluation DARPA et ils ont utilisé une valeur du facteur de pénalité égale à 1 qui est la valeur théorique standard de ce coefficient, mais dans la pratique elle n'est pas nécessairement égale à 1.

Plusieurs versions du BIC ont été développées, par la suite, pour effectuer de la segmentation. Parmi ces algorithmes, nous trouvons celui de (Delacourt, et al., 2000) qui a proposé une nouvelle méthode de segmentation qui consiste à faire deux passes : la première passe détecte les points de changements de locuteurs les plus probables, et la deuxième confirme ou infirme l'existence d'une frontière en utilisant une fenêtre d'analyse plus petite que celle utilisée pendant la première passe. Cette méthode possède l'avantage de détecter les changements de locuteurs qui sont proches l'un des autres.

Dans (Sivakumaran, et al., 2001) et (Cettolo, et al., 2005), la segmentation est basée sur le BIC et utilise une fenêtre d'analyse de taille variable en commençant avec une taille initiale et en l'augmentant tant qu'aucune frontière n'est détectée. Deux résolutions ont été proposées : une première consiste à calculer le ΔBIC avec une basse résolution. Si une frontière potentielle a été détectée pendant cette résolution, un calcul du ΔBIC est effectué à une haute résolution afin d'affiner la position du candidat détecté. Cette méthode permet d'avoir plus de

précision sur les positions des frontières des segments. Nous décrirons plus en détails le déroulement de cet algorithme dans le chapitre 4.

Enfin, il existe des méthodes de segmentation en tours de parole qui combinent les deux critères GLR et BIC (Meignier, et al., 2009), (El-Khoury, 2010). Une première passe de segmentation consiste à détecter les points de changement entre les locuteurs en se fondant sur le critère GLR, calculé en utilisant des gaussiennes avec des matrices de covariance pleines. Ensuite, une deuxième passe est effectuée avec le critère BIC pour fusionner les segments consécutifs d'un même locuteur.

2.4.2.4. Conclusion

Plusieurs paramètres acoustiques et critères de segmentation ont été proposés dans la littérature. Nous avons présenté les plus connus parmi eux pour la segmentation en tours de parole. Ces critères ont montré un pouvoir de discrimination entre locuteurs différents et de localisation de chaque début de tour de parole. Ce pouvoir nous a encouragés à tester ces méthodes pour segmenter en tours de chant.

2.4.3. Regroupement en locuteurs

Tandis que l'étape de segmentation opère sur les fenêtres adjacentes afin de déterminer si elles correspondent à un même locuteur ou non en détectant les points de changement entre locuteurs, l'étape de regroupement vise à identifier et regrouper ensemble les segments d'un même locuteur qui peuvent être localisés dans n'importe quelle région du flux audio. Dans un contexte d'indexation d'un enregistrement audio de dialogue entre plusieurs personnalités, la tâche de regroupement peut servir à regrouper les interventions par locuteur afin de mesurer, par exemple, dans les médias, pour chaque personnalité, son temps de parole et assurer le pluralisme. Cette tâche est aussi utilisée dans plusieurs applications telles que les systèmes de reconnaissance automatique de la parole qui utilisent des classes homogènes pour adapter les modèles acoustiques au locuteur avec l'algorithme MAP (*Maximum A Posteriori*), ce qui permet d'augmenter la performance de la reconnaissance.

Deux types de regroupement en locuteurs existent dans la littérature : regroupement à l'aveugle sans information *a priori* sur le nombre de locuteurs et leurs identités et regroupement avec information *a priori* sur le nombre de locuteurs et leurs identités. Dans notre étude, nous nous sommes intéressés au regroupement à l'aveugle en chanteurs ; ce problème peut être considéré comme un problème de classification non-supervisée. Les méthodes de classification non-supervisée utilisent généralement un regroupement hiérarchique.

- **Regroupement hiérarchique**

Le regroupement hiérarchique est une manière itérative de regrouper un ensemble d'éléments en fonction de l'objectif final et aussi en fonction du contenu. Les méthodes

existantes sont classées en deux catégories : les approches « bottom-up » et les approches « top-down », qui sont illustrées dans la Figure 2. 4. Les méthodes « bottom-up » sont initialisées avec un grand nombre de classes (chaque segment étant considéré comme une classe), alors que les méthodes « top-down » sont initialisées avec de très peu de classes ; généralement une seule classe est considérée et sera subdivisée à chaque itération. Pour les deux approches, l'objectif est de converger d'une manière itérative vers un nombre optimal de classes en utilisant soit un critère de fusion, pour la méthode « bottom-up », soit un critère de division, pour la méthode « top-down ». Le processus itératif se termine en utilisant un critère d'arrêt. Si le nombre de classes final est supérieur au nombre réel des locuteurs, alors il s'agit d'un sous-regroupement. Sinon, nous parlons d'un sur-regroupement.

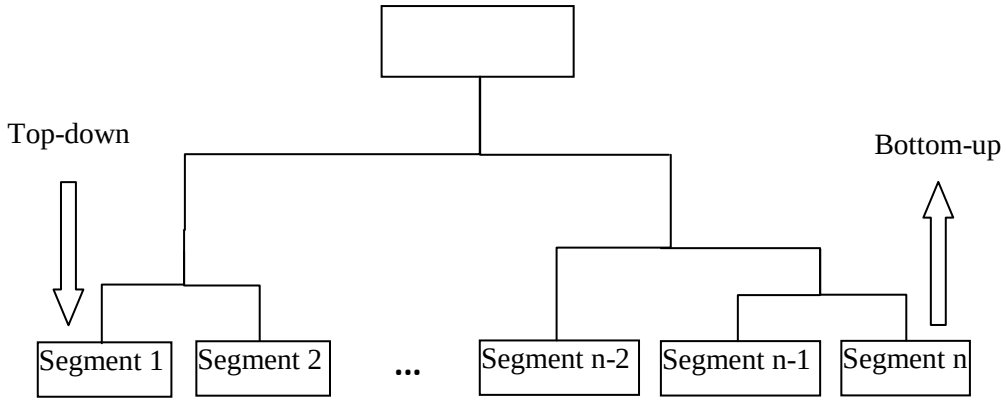

Figure 2. 4 – Les approches « bottom-up » et « top-down » du regroupement hiérarchique.

Les approches « bottom-up », connues aussi sous le nom de regroupement agglomératif, sont les plus utilisées dans la littérature puisqu'elles utilisent la sortie des processus de segmentation en locuteurs pour définir le point de départ du processus de regroupement. Ces systèmes de regroupement en locuteurs sont basés sur des métriques de fusion ou de découpage qui correspondent à une mesure de distance / similarité entre les classes.

Nous présentons dans les sections suivantes les approches existantes dans la littérature et les plus répandues en regroupement en locuteurs. Les méthodes exposées ci-dessous ont été initialement développées pour un traitement hors ligne (où nous pouvons accéder à tout l'enregistrement avant de le traiter). Mais certaines parmi elles peuvent être adaptées pour un traitement en ligne (où l'enregistrement complet n'est pas disponible).

2.4.3.1. Approches basées sur le BIC

Le Critère d'Information Bayésien (défini au paragraphe 2.4.2.3), proposé initialement par Chen et al. dans (Chen, et al., 1998), représente le critère de fusion le plus utilisé pour le regroupement en locuteurs. Il est utilisé de la façon suivante : à chaque itération, la valeur du ΔBIC est calculée pour toute paire de classe et la paire qui possède la valeur ΔBIC la plus faible est fusionnée. Ce processus est répété jusqu'à ce que toutes les paires aient une valeur

$\Delta BIC > 0$. En considérant deux classes C_1 et C_2, chacune est modélisée par une distribution gaussienne, la mesure ΔBIC,dans le cas des matrices de covariance pleines, est donnée par :

$$\Delta BIC = (n_1 + n_2)\log|\sum| - n_1 \log|\sum_1| - n_2 \log|\sum_2| - \frac{\lambda}{2}\left(d + \frac{d(d+1)}{2}\right)\log(n_1 + n_2) \quad (2.29)$$

où n_1, n_2 sont les tailles de C_1 et C_2. $\sum_1$, $\sum_2$ et $\sum$ sont respectivement les matrices de covariance de C_1, C_2 et $C_1 \cup C_2$. d est la dimension des vecteurs des paramètres. Dans le cas des matrices de covariance diagonales, la mesure ΔBIC est définie par :

$$\Delta BIC = (n_1 + n_2)\log|\sum| - n_1 \log|\sum_1| - n_2 \log|\sum_2| - \frac{\lambda}{2} d \log(n_1 + n_2) \quad (2.30)$$

2.4.3.2. Approche de modèle d'espace de vecteurs propres

L'approche de modèle d'espace de vecteurs propres (Eigen Vector Space Model – EVSM) proposée par Tsai (Tsai, et al., 2005) consiste à utiliser un modèle d'espace des vecteurs. Elle a été initialement proposée pour la recherche d'information afin de caractériser chaque mot comme un vecteur de termes acoustiques en se basant sur la méthode *tf-idf* (*Term Frequency-Inverse Document Frequency*), et ainsi avoir une mesure de similarité plus fiable entre les mots. La Figure 2. 5 décrit le déroulement de l'algorithme EVSM.

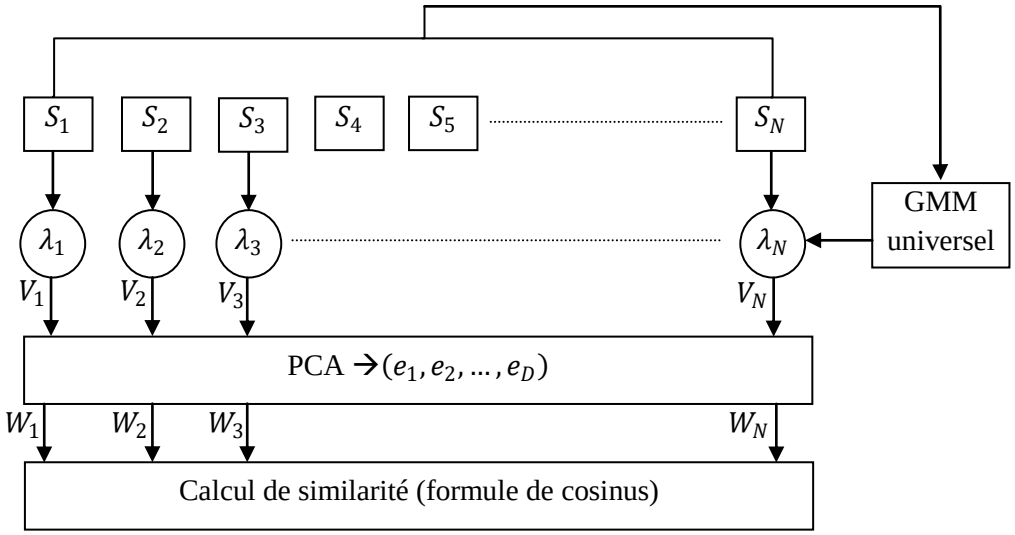

Figure 2. 5 – Illustration de l'algorithme d'EVSM.

La méthode EVSM consiste, tout d'abord, à apprendre un modèle du monde « GMM universel » à partir de tous les segments à regrouper. La méthode d'apprentissage est basée sur l'algorithme k-means (Wagsta, et al., 2001) pour l'étape d'initialisation suivie par l'application de l'algorithme Expectation-Maximisation (EM) (Dellaert, 2002). Ensuite, une adaptation du modèle du monde est effectuée pour chaque segment en utilisant l'estimation par Maximum *A Posteriori* (MAP) (Gauvain, et al., 1994).

Cette modélisation qui utilise le « GMM universel » permet d'avoir une estimation plus fiable des paramètres GMM pour les segments courts et d'instaurer une relation entre les composantes des GMM dérivés. Puis, un super-vecteur V_i est formé à partir de la concaténation de tous les vecteurs moyens de chaque GMM, en respectant cet ordre. Ensuite, une Analyse par Composante Principale (PCA) est appliquée sur l'ensemble des N super-vecteurs afin d'obtenir pour chaque super-vecteur V_i, un vecteur W_i de dimension réduite. Sont ainsi obtenus les D vecteurs propres, $e_1, e_2, \dots, e_D$, ordonnés par l'amplitude de leur contribution dans la matrice de covariance entre-mots :

$$B = \frac{1}{N}\sum_{i=1}^{N}(V_i - \bar{V})\,(V_i - \bar{V})' \tag{2.31}$$

avec $\bar{V}$ le vecteur moyen des V_i pour $1 \leq i \leq N$. Les D vecteurs propres constituent un espace propre, et chacun des super-vecteurs projetés dans cet espace peut être représenté par un point dans l'espace propre de coordonnées $\varphi_{i,d,d=1\dots D}$:

$$V_i = \bar{V} + \sum_{d=1}^{D}\varphi_{i,d}\,e_d \tag{2.32}$$

Enfin, la similarité est déterminée en utilisant la formule du cosinus :

$$S_{i,j}(V_i, V_j) = \frac{W_i \cdot W_j}{\|W_i\|\|W_j\|} \tag{2.33}$$

avec W_i et W_j correspondent aux vecteurs V_i et V_j obtenus après réduction de dimension (PCA).

2.4.3.3. Regroupement par CLR-CE

Plusieurs systèmes de segmentation et regroupement en locuteurs effectuent plus d'une étape de regroupement en utilisant un autre critère de fusion que celui utilisé pendant la première étape. Dans cette section, nous présentons les deux critères les plus répandus qui servent à la réalisation d'une étape finale de regroupement *a posteriori* dans plusieurs systèmes SRL. Le premier est le rapport de vraisemblance croisé (Cross Likelihood Ratio – CLR) (Reynolds, et al., 1998), (Zhu, et al., 2005) et le deuxième est l'entropie croisée (Cross Entropy – CE) (Le, et al., 2007), (Solomonoff, et al., 1998). Les deux scores CLR et CE donnent des résultats proches et le choix de l'un des deux dépend du corpus à traiter.

Après une première étape de regroupement (en utilisant le regroupement par BIC par exemple), la contribution du canal et les bruits de fond doivent être réduits et normalisés afin de permettre le regroupement des classes de locuteurs dont les conditions environnementales varient au cours de leur discours. De plus, la quantité de données présente dans chaque classe est suffisante pour permettre de construire un modèle de locuteur plus complexe et robuste de type GMM. Ainsi, un modèle du monde (Universel Background Model – UBM) est appris et

ensuite adapté pour chaque classe, en fournissant un modèle de locuteur correspondant à chacune des classes déjà obtenues.

Le score $CLR(i,j)$, calculé entre deux classes c_i et c_j contenant respectivement n_i et n_j données, est donné par l'expression suivante :

$$CLR(i,j) = \frac{1}{n_i}\log\frac{L(y_i|UBM)}{L(y_i|M_j)} + \frac{1}{n_j}\log\frac{L(y_j|UBM)}{L(y_j|M_i)} \tag{2.34}$$

où M_i et M_j correspondent respectivement aux modèles appris avec les données de la classe c_i et la classe c_j en adaptant l'UBM à l'ensemble des données y_i et y_j par l'algorithme MAP. $L(\cdot)$ est la fonction de vraisemblance. Pour le score CE, il est obtenu en remplaçant l'UBM dans les deux termes de l'équation (2.34) par les deux modèles M_i et M_j :

$$CE(i,j) = \frac{1}{n_i}\log\frac{L(y_i|M_i)}{L(y_i|M_j)} + \frac{1}{n_j}\log\frac{L(y_j|M_j)}{L(y_j|M_i)} \tag{2.35}$$

Le processus de décision avec ces deux scores est le même. A chaque itération, les classes qui minimisent le score $CLR(i,j)$ ou $CE(i,j)$ sont fusionnées : un nouveau modèle est appris pour la classe c_{i+j} et les scores entre cette nouvelle classe et toutes les autres sont calculés après chaque regroupement. Le processus s'arrête dès lors que toutes les paires possèdent une valeur de $CLR(i,j)$ ou $CE(i,j)$ supérieure à un seuil fixé *a priori*.

2.4.3.4. Regroupement par ILP / i-vecteurs

Les i-vecteurs ont été utilisés initialement dans le domaine de la vérification du locuteur (Dehak, et al., 2011). Ils permettent de réduire la quantité des données en vecteurs de dimension réduite, en ne gardant que la variabilité propre au locuteur et en enlevant les variabilités dues à des changements de conditions environnementales. Cette méthode a été adaptée par la suite à la segmentation et regroupement en locuteurs.

Dans les systèmes SRL, la méthode de regroupement par ILP / i-vecteurs est généralement utilisée comme étape finale de regroupement *a posteriori*, comme les deux scores CLR et CE décrits dans la section précédente. Le déroulement du processus de regroupement par cette approche se fait de la façon suivante : un i-vecteur est extrait pour chaque classe de locuteur en estimant tout d'abord un modèle de mélange de gaussiennes pour chaque classe par adaptation d'un modèle du monde aux données de la classe. Ensuite, un super-vecteur est formé par concaténation des moyennes des gaussiennes du modèle obtenu. Enfin, une projection de ce super-vecteur dans un espace de dimension réduite, optimisée de manière non-linéaire est effectuée afin de ne conserver que les informations pertinentes des locuteurs. Après extraction des i-vecteurs, une classification doit être appliquée afin de regrouper les i-vecteurs appartenant au même locuteur. La méthode de classification utilisée doit répondre à deux sous problèmes. Le premier consiste à minimiser le nombre de classes K

choisies parmi les N i-vecteurs et le deuxième consiste à minimiser la dispersion au sein des classes. La résolution de ces deux sous problèmes a été assimilée à la résolution d'un problème de programmation linéaire en nombres entiers (Integer Linear Programming – ILP) qui minimise l'équation (2.36) :

$$\sum_{k=1}^{N} x_{k,k} + \frac{1}{D}\sum_{k=1}^{N}\sum_{j=1}^{N} d(k,j)x_{k,j} \tag{2.36}$$

en vérifiant les contraintes suivantes :

$$x_{k,j} \in \{0,1\} \quad \forall\, k, \forall\, j \tag{2.37}$$

$$\sum_{k \in K_j} x_{k,j} = 1 \qquad \forall\, j\,, K_j = 1 \ldots N \tag{2.38}$$

$$d(k,j)x_{k,j} \leq \delta \quad \forall\, k, \forall\, j \tag{2.39}$$

$\sum_{k=1}^{N} x_{k,k}$ correspond au nombre des centres que nous cherchons à minimiser et $x_{k,k}$ est une variable binaire égale à 1 lorsque le i-vecteur k est un centre, 0 sinon. D est un facteur de normalisation et $x_{k,j}$ correspond à une variable binaire égale à 1 si le i-vecteur j est associé au centre k. Une distance $d(k,j)$ est calculée entre le centre k et le i-vecteur j en utilisant la distance de Mahalanobis (Bousquet, et al., 2011):

$$d(k,j) = (k-j)\sum^{-1}(k-j)' \tag{2.40}$$

avec $\sum$ la matrice de covariance intra-classe calculée sur les i-vecteurs. Pour tout i-vecteur j associé au centre k, la distance $d(k,j)$doit être inférieure à un seuil δ fixé *a priori*.

2.4.3.5. Autres techniques de regroupement

Il existe aussi d'autres méthodes de regroupement dans la littérature telles que l'approche de regroupement par Modèles de Markov Cachés qui consiste à considérer chaque classe comme un état du modèle et les changements entre les locuteurs sont caractérisés par les transitions entre les états. Le système de regroupement décrit dans (Ajmera, et al., 2002) est basé sur un HMM ergodique avec des contraintes de durée minimale et un sur-regroupement de données effectué pendant la première étape. Ensuite, un apprentissage des HMM est réalisé avec l'algorithme EM. Le log du rapport de vraisemblance (Log Likelihood Ratio – LLR) est utilisé comme critère de fusion entre deux classes.

Le système SRL décrit dans (Meignier, et al., 2001) est effectué en une seule passe combinant les deux étapes de segmentation et regroupement, ce qui n'est pas le cas des autres systèmes SRL. Ce système considère, au départ, qu'un seul locuteur est présent sur la totalité de l'enregistrement de la parole, caractérisé par un seul état. Ensuite, il se base sur un processus itératif qui permet de générer les HMM, détecter les nouveaux états (les nouveaux locuteurs) et les ajouter à la configuration des HMM. Le critère d'arrêt utilisé est basé sur une

comparaison de la probabilité tout au long du chemin de Viterbi entre deux itérations du processus.

D'autres approches de regroupement en locuteurs procèdent d'une manière différente des méthodes décrites ci-dessus, en définissant des métriques pour déterminer le nombre de classes *a priori* et trouver par la suite le regroupement optimal qui permet d'avoir ce nombre de classes.

Parmi ces approches, nous trouvons celle décrite dans (Tsai, et al., 2007) dans laquelle le nombre de locuteurs est calculé en utilisant le BIC. Le regroupement optimal qui optimise la vraisemblance globale du modèle est déterminé en utilisant un algorithme génétique. Des cartes d'auto-organisation (self-organizing) sont proposées dans (Lapidot, 2003) pour le regroupement en locuteurs en utilisant un algorithme de Quantification Vectorielle (QV) pour former des *code-books* représentant chacun des locuteurs.

2.4.4. Les systèmes « complets » de segmentation et regroupement en locuteurs

Nous nous intéressons dans cette section à l'étude du déroulement d'un système « complet » de segmentation et regroupement en locuteurs. L'architecture générale des systèmes SRL les plus connus est illustrée dans la Figure 2. 6. Elle est composée de cinq étapes principales : la première correspond à une extraction des paramètres acoustiques, la deuxième consiste à faire la détection d'activité vocale pour détecter les régions de la parole, la troisième est la segmentation en locuteur en utilisant l'approche combinée GLR/BIC et les deux dernières étapes correspondent au regroupement en locuteur en utilisant tout d'abord le BIC et ensuite le CLR ou le CE. Dans cette partie, nous présentons deux exemples de systèmes SRL qui ont apporté quelques modifications à cette architecture standard : le système de l'IRIT et le système du LIUM.

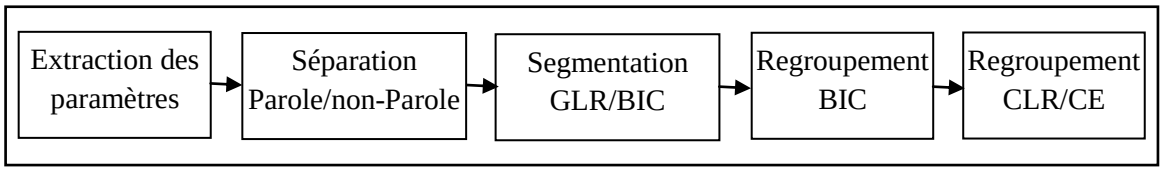

Figure 2. 6 – Architecture standard d'un système de segmentation et regroupement en locuteurs.

2.4.4.1. Système de segmentation et regroupement en locuteurs de l'IRIT

Dans (El-Khoury, 2010), les auteurs proposent des améliorations pour la détection de l'activité vocale, la segmentation en locuteurs et pour le regroupement en locuteurs en utilisant un nouveau schéma itératif qui considère ces trois tâches comme un seul problème. Le déroulement de l'algorithme est illustré dans la Figure 2. 7.

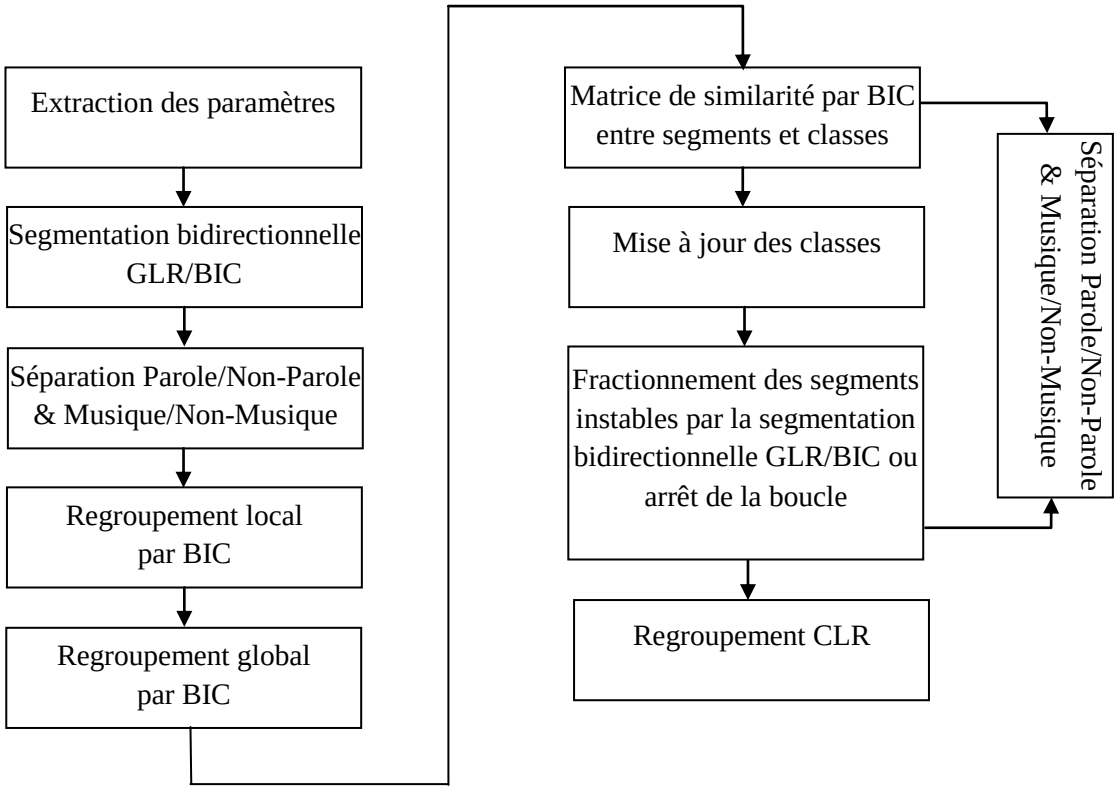

Figure 2. 7 – Architecture du système de segmentation et regroupement en locuteurs de l'IRIT.

L'architecture de ce système est composée de onze étapes :

1. Extraction des paramètres MFCC, la modulation d'énergie à 4Hz et la vraisemblance de la trame d'analyse par rapport à deux GMM (un caractéristique de la parole et un de la non-parole).

2. Segmentation bidirectionnelle avec le GLR/BIC consiste à appliquer le même traitement en procédant de « gauche à droite » et de « droite à gauche » du signal parce qu'il y a des chances qu'une frontière ratée dans la première direction puisse être trouvée dans la deuxième direction et vise versa. La valeur du coefficient de pénalité λ_1 est pratiquement égale à 1.

3. Séparation en parole/non-parole en utilisant la modulation d'énergie à 4 Hz et les scores des GMM de la parole et non-parole pour chaque segment est appliquée.

4. Séparation en musique/non-musique en utilisant le nombre et la durée des segments ainsi que les scores des GMM de la musique et non musique.

5. Regroupement local par BIC consiste à effectuer un regroupement tous les N segments consécutifs avant de faire le regroupement global et cela permet d'initialiser les classes

avec des données plus importantes pour éviter de faire une comparaison entre des classes de tailles réduites.

6. Regroupement global par BIC basé sur les classes obtenues de l'étape précédente.

7. Calcul de la matrice de similarité entre les segments $S_i (i = 1, \dots, N_s)$ et les classes $C_j (j = 1, \dots, N_c)$ où N_s est le nombre de segments et N_c est le nombre de classes.

8. Mise à jour des classes en attribuant chaque segment S_i à $arg\max_{C_j} \left(-\Delta BIC(S_i, C_j)\right)$ avec j variant de 1 à N_c.

9. Fractionnement des segments instables en utilisant la segmentation bidirectionnelle GLR/BIC avec une valeur du coefficient de pénalité λ_2 égale à 0,8. Un segment instable correspond à un segment dont la similarité avec sa classe correspondante est faible.

10. Arrêt de la boucle s'il n'y a plus de fractionnement à faire. Sinon, les étapes de 3 jusqu'à 7 sont à refaire.

11. Regroupement final par CLR pour regrouper les classes correspondantes au même locuteur mais avec des conditions environnementales différentes.

Ce système a donné des bonnes performances par rapport au système SRL standard. En effet, un taux d'erreur de 11,01% a été obtenu avec ce système, lorsqu'il a été appliqué sur le corpus ESTER-2, alors qu'avec le système standard dont l'architecture est illustrée dans la figure 1.6, un taux d'erreur de 17,42% a été trouvé sur ce même corpus.

2.4.4.2. Système de segmentation et regroupement en locuteurs du LIUM

Le système du LIUM (Meignier, et al., 2009) est parmi les systèmes de segmentation et regroupement en locuteurs les plus connus ; il est d'ailleurs disponible en ligne[2]. Son architecture est présentée dans la Figure 2. 8 et il comprend six étapes que nous énumérons :

1. Extraction des paramètres acoustiques MFCC.

2. Segmentation par BIC qui consiste, dans une première étape, à découper le signal en des segments de même taille, généralement égale à 2,5 secondes. Ensuite, les segments consécutifs qui sont acoustiquement homogènes sont fusionnés en utilisant la distance BIC.

3. Regroupement par BIC des segments qui ont été prononcés par le même locuteur dans une seule classe.

[2] http://www-lium.univ-lemans.fr/diarization/doku.php/overview

4. Re-segmentation par Viterbi pour affiner les frontières des segments obtenus à partir de la deuxième étape. Chaque classe est modélisée par un HMM à un seul état, représenté par un GMM de 8 gaussiennes appris sur tous les segments de la classe. Le seuil du log de la pénalité entre deux HMM est fixé expérimentalement.

5. Détection de la parole qui consiste à effectuer une segmentation en parole/non-parole réalisée en utilisant un décodage de Viterbi avec un HMM à 8 états dont deux états correspondent au silence (bande large et étroite), 3 à de la parole à large bande (propre, en présence de bruit de fond, en présence de musique de fond), 1 à de la parole à bande étroite, 1 pour les jingles et 1 pour la musique.

6. Regroupement par CLR ou par CE pour regrouper toutes les classes liées à un seul locuteur dans une seule classe.

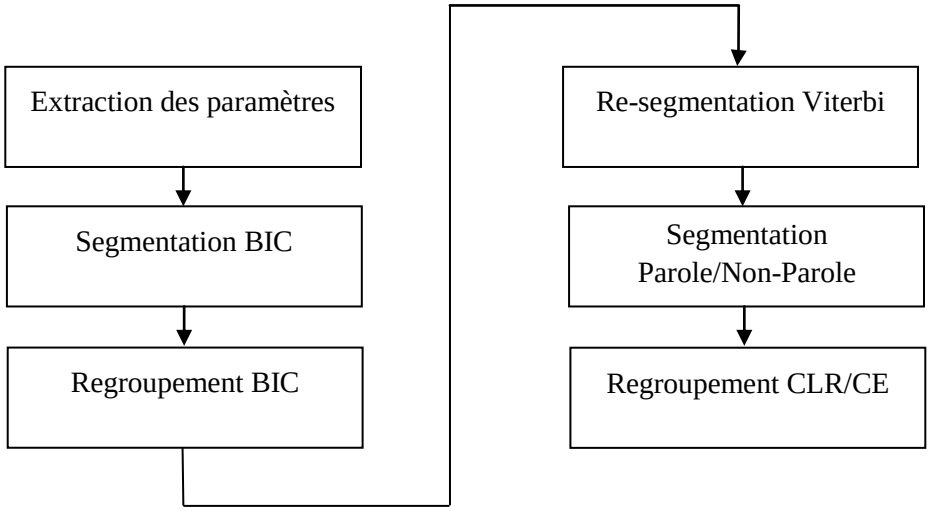

Figure 2. 8 – Architecture du système de segmentation et regroupement en locuteurs du LIUM.

L'application de ce système sur le corpus ESTER-2 en utilisant la distance CLR pour la dernière étape du regroupement a donné un taux d'erreur de 11,27%, très proche du résultat trouvé avec le système SRL de l'IRIT.

2.5. Conclusion

Dans ce chapitre, nous avons présenté une brève revue des méthodes et outils utilisés pour la segmentation et le regroupement en locuteurs en vue de réaliser un système dédié au chant. Nous nous sommes focalisés tout d'abord sur les travaux réalisés en traitement du chant au sens large ainsi que sur les paramètres utilisés pour caractériser ce type de signal sonore. Puis, nous avons étudié plus en détails les approches utilisées en segmentation et regroupement en locuteurs.

Plusieurs méthodes ont été développées en traitement automatique du chant (détection de la voix chantée, segmentation solo / chœur) et de nombreux paramètres acoustiques ont été proposés : paramètres classiques du traitement audio (ZCR, énergie, centroïde spectral, flux spectral), paramètres spécifiques à la musique et au chant (coefficient harmonique, timbre, vibrato, tempo) ainsi que des paramètres issus de la modélisation tels que les chroma.

La segmentation et regroupement en locuteurs est bien avancée depuis des années et plusieurs paramètres et différentes approches pour la segmentation et le regroupement ont été proposées. Les méthodes de paramétrisation pour cette tâche sont basées généralement sur la modélisation (FBANK, MFCC, PLP, RASTA-PLP) et permettent de prendre en compte, à partir des échelles non-linéaires (Mel, Bark), la réponse du système auditif humain. La tâche de segmentation est réalisée en utilisant des métriques de décision (KL2, GLR, BIC) qui permettent de confirmer ou non l'existence d'un point de changement de locuteur. En ce qui concerne le regroupement en locuteurs, la plupart des approches réalisent un regroupement hiérarchique agglomératif (EVSM, BIC, CLR, CE, ILP / i-vecteurs). Ces paramètres, critères et approches utilisés pour la segmentation et le regroupement ont permis d'avoir de bonnes performances pour les systèmes SRL.

Le pouvoir de discrimination entre locuteurs lors de la segmentation et la capacité des méthodes développées pour regrouper les segments d'un même locuteur impliquent la possibilité d'appliquer ces techniques à des enregistrements audio autres que la parole. Ainsi, elles peuvent être utilisées pour effectuer de la segmentation et regroupement en chanteurs car le chant est comme la parole produit par le même instrument (la voix humaine). Néanmoins, certaines caractéristiques du chant qui le rapprochent de la musique par son contexte d'utilisation (à l'instar des instruments), peuvent produire de difficultés avec les techniques utilisées pour la parole.

Chapitre 3

Définitions, corpus et annotation

Sommaire

3.1. Introduction

Dans notre contexte de segmentation et regroupement en chanteurs, il convient tout d'abord de définir un nouveau concept qui est le concept « **tour de chant** » ainsi que les situations de changement qui impliquent l'existence ou non d'un nouveau tour de chant. Nous consacrons ce chapitre pour définir ce concept et présenter notre contexte de validation, qui est utilisé pour la validation de nos systèmes de segmentation et regroupement avant de les appliquer sur le corpus du projet DIADEMS qui est le contexte applicatif des travaux de cette thèse. Ce dernier est fourni par les ethnomusicologues du projet et composé des enregistrements hétérogènes, qui sont détaillés dans le chapitre 6. Pour la mise en place et l'évaluation des outils développés pendant la thèse, il nous a semblé important de constituer un autre corpus de travail, composé d'enregistrements plus homogènes et de qualité de type « studio d'enregistrement ». Dans ce chapitre, notre étude est faite sur ce corpus de validation.

La définition du terme « **tour de chant** » et les situations de changement sont présentées dans la première section de ce chapitre. La deuxième section est consacrée à la description du corpus utilisé pour évaluer et valider la robustesse de notre système de segmentation et regroupement en chanteurs. Les différentes règles et conventions utilisées pour annoter en tours de chant et classes de chanteurs font l'objet d'une troisième section. Dans une dernière section, nous présentons les critères d'évaluation utilisés pour les deux étapes de segmentation et regroupement de notre système.

3.2. Définition d'un tour de chant

A notre connaissance, dans la littérature, il n'y a ni de définition du tour de chant, ni de travail effectué sur les tours de chant. En revanche, en traitement automatique de la parole, il y a de nombreuses études sur les tours de parole (segmentation et regroupement en locuteurs). Comme le chant possède des caractéristiques qui le rapproche de la parole par son mode de production (tous les deux sont produits par la voix humaine), nous nous sommes inspirés des définitions existantes sur les tours de parole afin d'essayer de définir les tours de chant. Ainsi, parmi les définitions existantes, citons celle-ci: *« Par tour de parole, on entend le mécanisme d'alternance des prises de parole. L'unité qui constitue cette alternance est la contribution verbale d'un locuteur à un moment déterminé de l'échange »* (UCL, 2002). Cette définition considère un tour de parole comme un mécanisme d'alternance entre différents locuteurs qui parlent à tour de rôle.

Dans le cadre du projet DIADEMS, les ethnomusicologues ont essayé de définir le terme « **tour de chant** » en proposant trois définitions. La première considère qu'un tour de chant est une *« Alternance entre deux ou plusieurs groupes de chanteurs »*. La deuxième le définit comme une *« Alternance entre différents groupes de chanteurs »*. La troisième proposition considère ce terme comme une *« Alternance de différents intervenants (soliste(s), chœur(s)) présentant des parties similaires ou distinctes au niveau du contenu mélodico-rythmique et de l'énoncé linguistique »*. Ces trois propositions associent la notion de tour de chant à une

alternance entre différents chanteurs quelque soit le contenu musical (mélodie, rythme) et linguistique (paroles) de leurs parties chantées.

Suite à ces différentes définitions, nous avons retenu que la segmentation en tour de chant est associée à un mécanisme d'alternance entre différents groupes de chanteur(s) quelque soit le contenu du morceau chanté. Nous précisons qu'un groupe de chanteurs peut être composé d'un seul chanteur (soliste) ou de plusieurs (chœurs).

De ces considérations, nous proposons la définition suivante :

> **La segmentation en tour de chant** est un mécanisme d'*alternance* entre différents groupes de chanteur(s) que les parties soient similaires ou distinctes au niveau du contenu mélodico-rythmique et de l'énoncé linguistique.

Après avoir défini le mécanisme de « segmentation en tour de chant », il convient de préciser les situations de changement c'est-à-dire les moments d'alternance où nous devons décider s'il s'agit d'un nouveau groupe de chanteur(s) (nouveau tour de chant) ou non. Un changement est considéré dans les cas suivants :

- Passage d'un groupe de i chanteurs $G_i (i = 1 \dots N)$ à un groupe de j chanteurs $G_j (j = 1 \dots N')$, tel que les deux groupes diffèrent d'au moins un chanteur :
$$G_i \neq G_j \ \forall \ i, j$$
- Passage d'un groupe de chanteur(s) à une zone de non chant (silence, parole, instrument...) et vice versa.

La Figure 3. 1 illustre ces deux changements de situation de tour de chants. La première alternance entre le Tour_k et le Tour_{k+1} représente la première situation de changement. La deuxième alternance entre Tour_{k+1} et la zone de Silence, et la troisième alternance entre le Silence et le Tour_{k+2}, représentent la deuxième situation du changement.

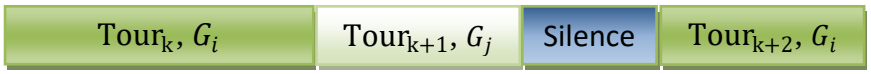

Figure 3. 1 – Situations de changement de tour de chants.

3.3. Corpus

Dans notre travail, nous utilisons deux corpus différents : le premier est appelé « corpus studio » et le second « corpus DIADEMS ». Dans ce chapitre, nous ne décrivons que le premier corpus. En effet, le second sera décrit lors des expérimentations dédiées au projet DIADEMS (chapitre 6), avec la description des usages associés. A noter que les résultats présentés dans les quatrième et cinquième chapitres sont obtenus sur le corpus « studio ».

3.3.1. Description générale du corpus

Les enregistrements du corpus « studio » ont été effectués dans des conditions acoustiques contrôlées et contiennent seulement du chant. Ils ont été extraits à partir de quelques albums de musique qui contiennent du chant sans instrument, à savoir la chanson intitulée « Mayingo » de l'album « Lambarena Bach to Africa », les pistes vocales de la chanson « Sloop John B » des Beach Boys et la chanson « Marions les Roses » du groupe « Malicorne ».

Ce corpus est constitué de 9 enregistrements d'une durée totale d'environ 11 minutes ; nous l'avons divisé en un ensemble de développement (DEV) et un ensemble d'évaluation (EVAL) dans les proportions 31% et 69% respectivement. Le Tableau 3. 1 décrit cette répartition du corpus. Le DEV comporte 10 groupes de chanteur(s) : il est utilisé pour l'ajustement des paramètres acoustiques et des hyper-paramètres des algorithmes de segmentation et de regroupement (cf. Chapitre 4 et Chapitre 5). Afin de valider la performance des systèmes de segmentation et regroupement en chanteurs, nous utilisons l'ensemble EVAL qui comporte 13 groupes de chanteur(s).

Tableau 3. 1 – Répartition et description du corpus « studio ».

Ensembles	**DEV**	**EVAL**
Nombre d'enregistrements	4	5
Durée	196 secondes	434 secondes
Nombre de groupes de chanteurs	10	13

Les caractéristiques techniques des fichiers sonores utilisés sont : 16 bits, 16 kHz et mono.

3.3.2. Différentes situations de changement rencontrées dans le corpus

Ce corpus contient des catégories différentes d'alternance entre groupes de chanteur(s), en fonction des identités des chanteurs présents dans le groupe.

- La première catégorie correspond aux alternances entre deux solistes différents :

$Tour_k$, $Soliste_i$	$Tour_{k+1}$, $Soliste_j$	$Tour_{k+2}$, $Soliste_i$

- La deuxième comprend les alternances entre deux chœurs différents :

$Tour_k$, $Chœur_i$	$Tour_{k+1}$, $Chœur_j$	$Tour_{k+2}$, $Chœur_i$

- La troisième est représentée par les alternances entre un soliste et un chœur :

$Tour_k$, $Soliste_i$	$Tour_{k+1}$, $Chœur_j$	$Tour_{k+2}$, $Soliste_i$

- La quatrième contient les alternances entre un soliste ou un chœur et du silence :

$Tour_k$, G_i	Silence	$Tour_{k+1}$, G_i

Pour résumer, les trois premières catégories d'alternances rencontrées dans le corpus « studio » correspondent à la première situation de changement décrite dans la section 3.2. , à savoir un passage d'un groupe de chanteur(s) à un autre. La quatrième catégorie est incluse dans la deuxième situation de changement qui correspond au passage d'une zone de chant à une zone de non chant. Ces deux situations sont représentées à parts égales dans le corpus. En effet, nous trouvons en général autant de cas de la première situation que la deuxième situation.

3.4. Annotation

Afin de pouvoir évaluer un système de segmentation et regroupement en chanteurs, nous avons besoin d'établir une vérité terrain, c'est-à-dire une référence. Pour cela, nous avons annoté manuellement nos corpus en des segments de tours de chant et des classes de groupes de chanteurs. Dans cette partie, nous parlons tout d'abord des conditions d'annotation, ensuite, nous présentons les conventions que nous avons mises en place pour assurer un bon processus d'annotation et leur mise en œuvre.

3.4.1. Conditions d'annotation

L'annotation s'effectue idéalement avec un casque, dans un environnement sans bruit et seulement avec la composante sonore d'un enregistrement. L'annotateur dispose de la forme d'onde et du spectrogramme du signal audio. Nous avons utilisé le logiciel « Sonic Visualiser »[3]. La Figure 3. 2 montre l'interface de ce logiciel et l'affichage du spectrogramme d'un extrait de chant du corpus « studio ».

A l'aide de cette interface, un étiquetage des différentes zones est possible. Sur la Figure 3. 2, le segment bleu entouré en rouge représente l'insertion d'une étiquette mentionnant que ce morceau a été chanté par le « Soliste 1 ». Cette insertion implique la présence de deux frontières de début et de fin du segment désignant les points de changement entre chanteurs et par conséquent les instants de début et de fin du tour de chant correspondant au « Soliste 1 ». Le fichier annoté est exporté au format XML ou texte. Le fichier d'annotation exporté comporte le début, la fin ou la durée de chaque segment en secondes ou en trames selon le choix de l'utilisateur ainsi que l'étiquette correspondante à chaque segment.

[3]http://www.sonicvisualiser.org/

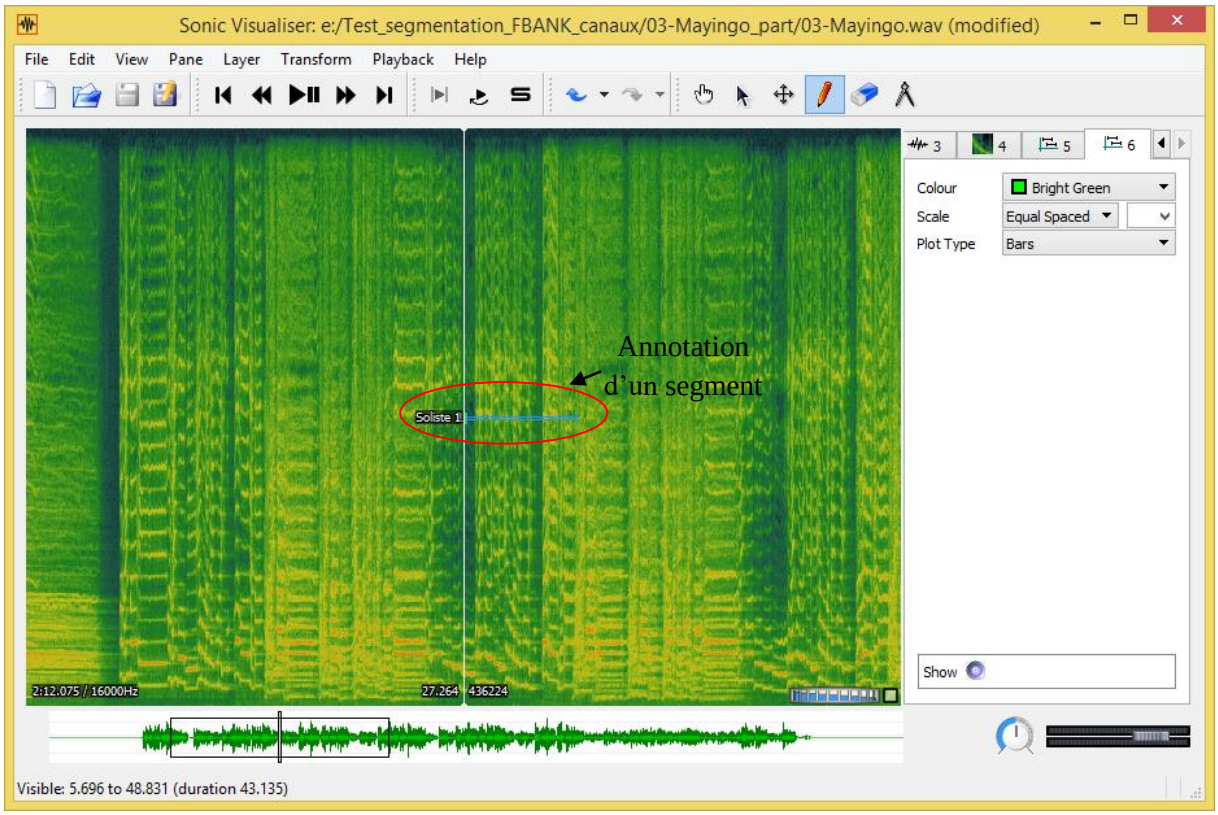

Figure 3. 2 – Exemple de visualisation d'un segment en « Soliste 1 » avec le logiciel « Sonic Visualiser ». Il s'agit d'un extrait de 43 secondes du fichier « 03-Mayingo ».

3.4.2. Conventions d'annotation

Nous présentons dans cette partie les règles d'annotation que nous avons élaborées pour pouvoir assurer un déroulement fiable du processus d'annotation. Par la suite, nos travaux se divisent en deux étapes principales qui sont la segmentation en tours de chant et le regroupement en chanteurs, processus que nous avons voulu évaluer séparément. De manière comparable, nous avons mis en place, tout d'abord, les conventions d'annotation pour segmenter en tours de chant en se basant sur notre définition du terme tour de chant, sur les deux situations de changement déterminées et les conventions d'annotation utilisées pour les tours de parole (Ester2, 2008). Ensuite, nous avons défini les règles d'annotation pour regrouper en groupe de chanteurs. La Figure 3. 3 montre un extrait d'un enregistrement du corpus « studio » annoté en tours de chant, c'est-à-dire segmenté en tours de chant et la Figure 3. 4 illustre le même extrait, mais après annotation en classes de chanteurs, c'est-à-dire après regroupement en groupe de chanteurs.

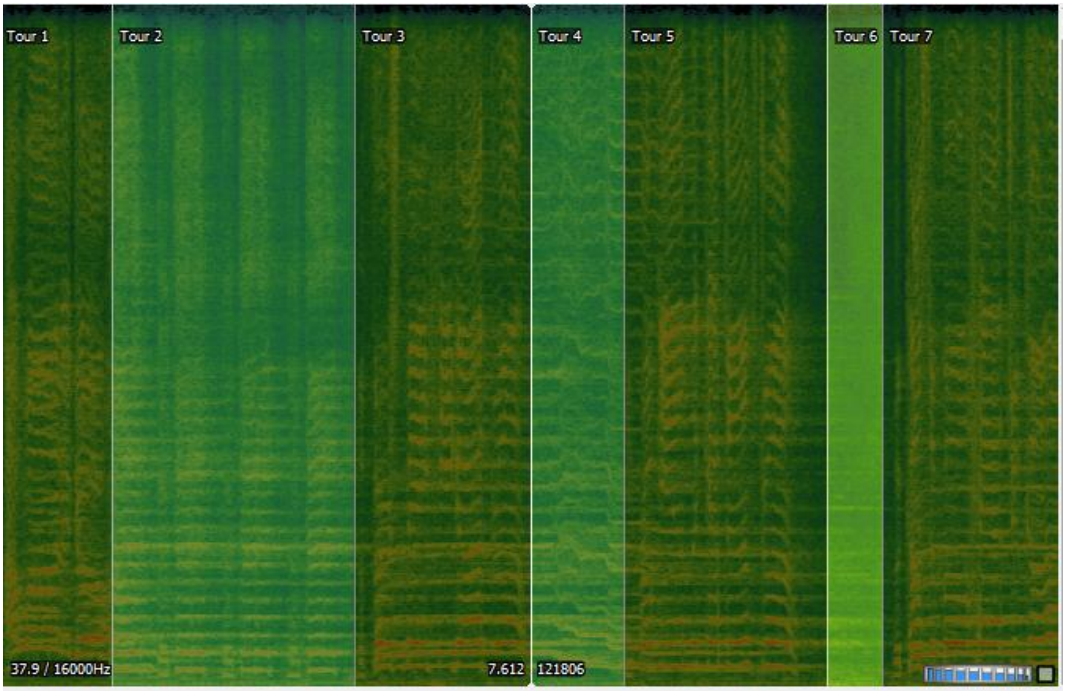

Figure 3. 3 – Exemple d'un extrait d'un enregistrement du corpus « studio » annoté en tours de chant. Il s'agit d'un extrait de 15 secondes du fichier « sloopJohnB_dev ».

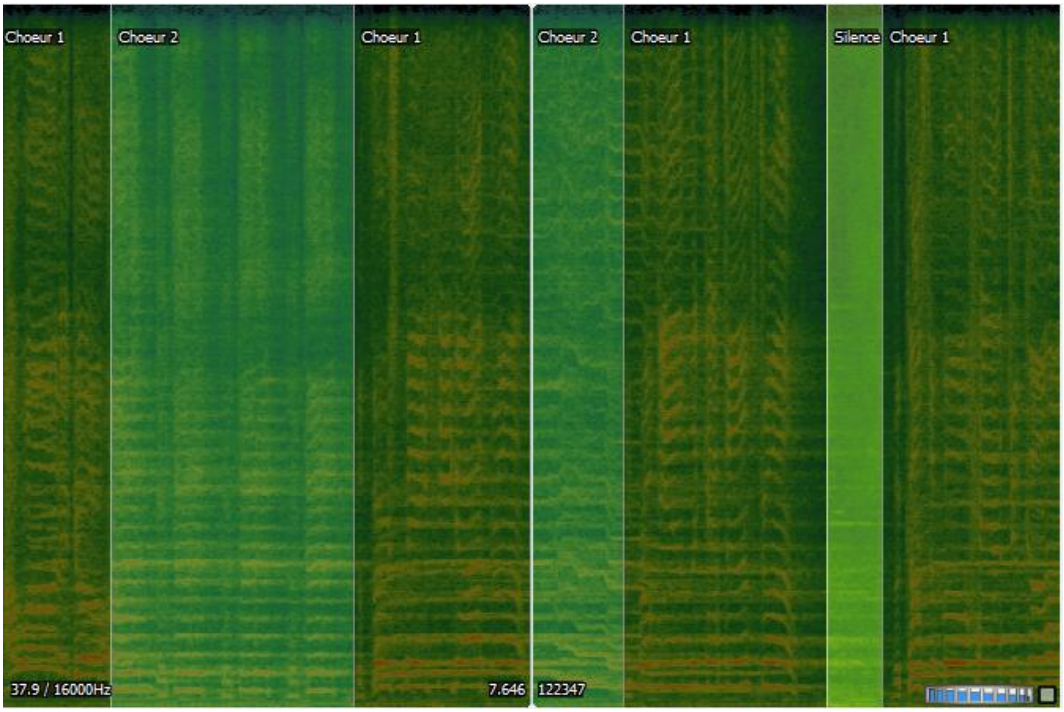

Figure 3. 4 – Exemple d'un extrait d'un enregistrement du corpus « studio » annoté en classes de chanteurs. Il s'agit du même extrait de 15 secondes du fichier « sloopJohnB_dev » visualisé sur la Figure 3. 3.

Les conventions d'annotation en tours de chant sont présentées dans les cinq premières sous-parties de cette section. La dernière sous-partie illustre le déroulement de l'annotation pour regrouper en chanteurs.

3.4.2.1. Segments de chant : frontières et durée

Pour décider s'il s'agit d'un tour de chant ou non, nous avons consulté quelques guides d'annotation existants pour les tours de parole (DGA, 2008), (RT03, 2003), (RT09, 2009), (Gravier, et al., 2012). Ces guides préconisent de caractériser un segment par son début et sa durée ; ils considèrent que les frontières des tours représentent le début de l'intervention du locuteur et que la durée de chaque tour ne doit pas être trop courte : la durée d'un segment de parole doit être supérieure à 0,3 ou 0,5 secondes (RT09, 2009). A partir de ces deux règles, nous avons déduit, tout d'abord, une première convention qui consiste à ne considérer comme frontières des segments de chant que le début de chaque tour de groupe de chanteur(s). Ensuite, nous avons essayé de mettre une deuxième convention pour définir la durée minimale d'un tour de chant. En annotant quelques fichiers, nous nous sommes aperçus qu'il y a parfois des alternances rapides qui impliquent l'existence des tours de durée courte. Donc, nous avons décidé de ne pas mettre de contrainte sur la durée minimale d'un tour de chant.

3.4.2.2. Nouveau segment de chant

Un nouveau tour de chant est indiqué par l'insertion d'une frontière au début de chaque changement de groupe de chanteur(s). Si la fin d'un tour de chant est suivie d'un début d'un autre tour, nous n'avons pas besoin de spécifier une frontière pour indiquer la fin du tour précédent. Si le tour de chant est suivi d'une pause longue de non chant (parole, silence, instruments...), les annotateurs doivent rajouter une frontière au début de la pause que nous considérons comme un tour de non-chant. Il convient alors de préciser les longueurs (durées) des segments de non chant. Par conséquent, le fichier de sortie contient le début de chaque tour de chant ou de non-chant ainsi que sa durée.

3.4.2.3. Courte et longue périodes de non chant

Lors de notre étude, nous avons remarqué qu'un groupe de chanteur(s) ne chante pas tout le temps, il fait souvent des pauses. Dans des enregistrements musicaux, il y a ainsi des pauses longues, qui peuvent être des silences ou des transitions instrumentales. Nous avons décidé de considérer qu'une pause est longue si sa durée est supérieure à un seuil D_{min}. Elle est alors identifiée comme un tour de non-chant. La valeur de D_{min} a été fixée à 0,5 secondes. Par contre, les pauses courtes (inférieure à D_{min}) ne sont pas annotées : elles sont incluses dans le tour de chant précédent.

Lors de notre annotation en tours de chant en utilisant les trois conventions décrites ci-dessus, nous avons rencontré des cas idéaux. La Figure 3. 5 est un extrait du DEV du corpus « studio » qui est composé d'alternances entre groupes de chanteur(s) dont la détermination des débuts et fins de chaque tour de chant est facile et rapide. Mais, comme il existe des cas simples à annoter, il existe aussi certains cas difficiles. Parmi ceux-ci, nous trouvons les cas de courte superposition entre groupe de chanteur(s) et les cas des alternances rapides. Pour cela, nous avons défini des conventions pour ces cas de courte superposition et d'alternance rapide pour rendre le processus d'annotation plus objectif et donc fiable dans ces situations. Ces conventions sont illustrées dans les deux paragraphes suivants.

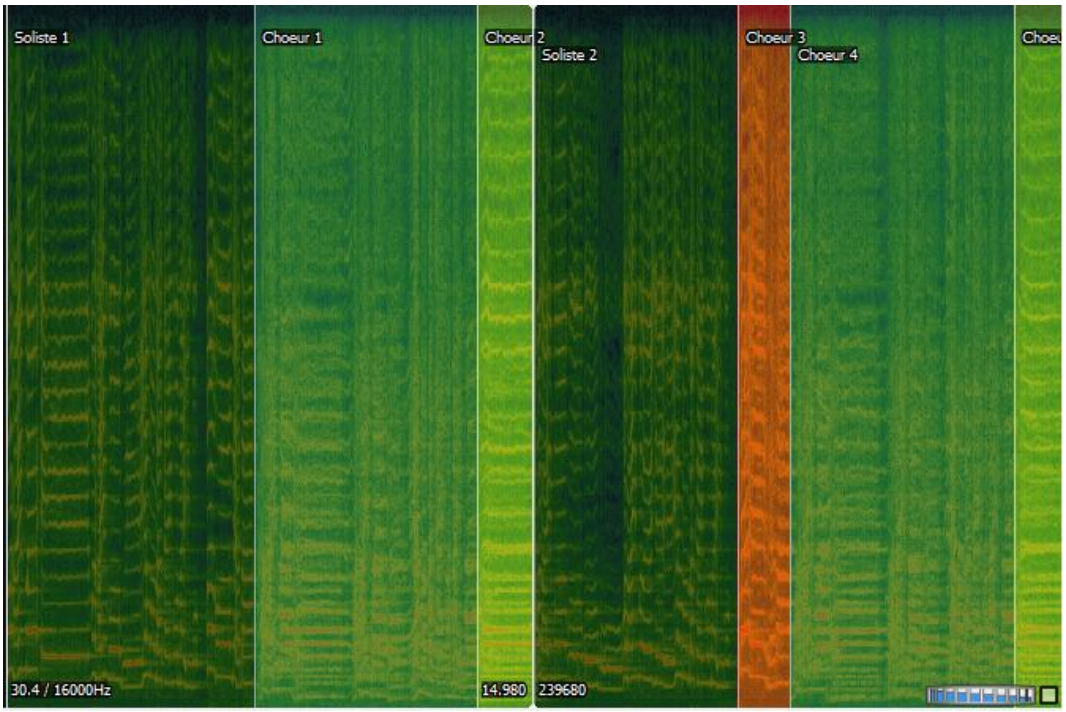

Figure 3. 5 – Illustration d'une annotation manuelle en tours de chant d'un extrait de 30 secondes du fichier « 03-Mayingo_dev ».

3.4.2.4. Courte superposition entre groupe de chanteur(s)

Comme pour la parole, nous pouvons rencontrer des régions de superposition entre groupes de chanteur(s). Ces régions de superposition, dont la durée est supérieure à 0,5 seconde, sont annotées en tant que nouveau tour de chant composé d'un nouveau groupe de chanteurs, qui est l'union des deux groupes superposés. Néanmoins, il y a des cas où la durée de la superposition entre les groupes de chanteurs est inférieure à 0,5 seconde tels que le cas présenté dans la Figure 3. 6.

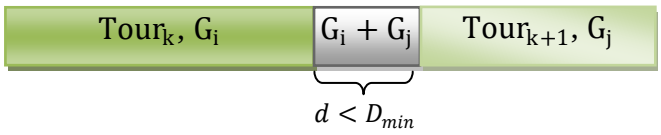

Figure 3. 6 – Cas de courte superposition entre groupe de chanteur(s).

Dans ces cas, il s'est avéré impossible d'introduire une valeur minimale pour décider si nous considérons la superposition comme un nouveau tour de chant ou non. Par conséquent, nous avons décidé de considérer la partie de superposition entre les deux groupes de chanteur(s) comme un nouveau tour de chant composé de l'ensemble des chanteurs du $\mathrm{Tour_k}$ et $\mathrm{Tour_{k+1}}$, quelque soit la durée de la superposition.

3.4.2.5. Alternance rapide

Dans quelques régions de nos enregistrements, nous avons des alternances rapides entre les différents groupes de chanteur(s), dont la durée ne dépasse pas 0,5 secondes. La Figure 3. 7 illustre ces cas.

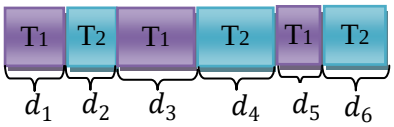

Figure 3. 7 – Illustration d'un cas limite d'alternances rapides.

Avec $d_i < 0.5\ s, _{i=1\ldots6}$

Dans ce cas, nous avons éprouvé des difficultés à placer les frontières des segments aux bons endroits (instants de changements entre les groupes de chanteurs). La Figure 3. 8 montre le spectrogramme et les frontières d'un segment de tour de chant pour un morceau qui représente un cas d'alternance rapide de durée 0,4 secondes (segment en rouge). Néanmoins, pour certains cas, il y a des alternances qui sont tellement rapides que nous n'avons pas pu définir les points de début et fin des tours qui les composent. Dans ces cas, nous avons décidé de considérer toute la partie des alternances très rapides comme un seul tour de chant. Nous avons défini un seuil D_{inf} qui est la durée nécessaire pour retenir une alternance très rapide ; au-dessous de cette durée, on ne découpe pas et on considère ces alternances comme un seul tour. D_{inf} est égal à 0,3 s. La Figure 3. 9 illustre un cas d'alternance très rapide pour lequel il s'est avéré impossible de déterminer les points de changement entre les groupes de chanteurs. Le segment en vert représente le tour de chant constitué de plusieurs alternances très rapides : plus que 25 alternances de durées inférieures à 0,3 s.

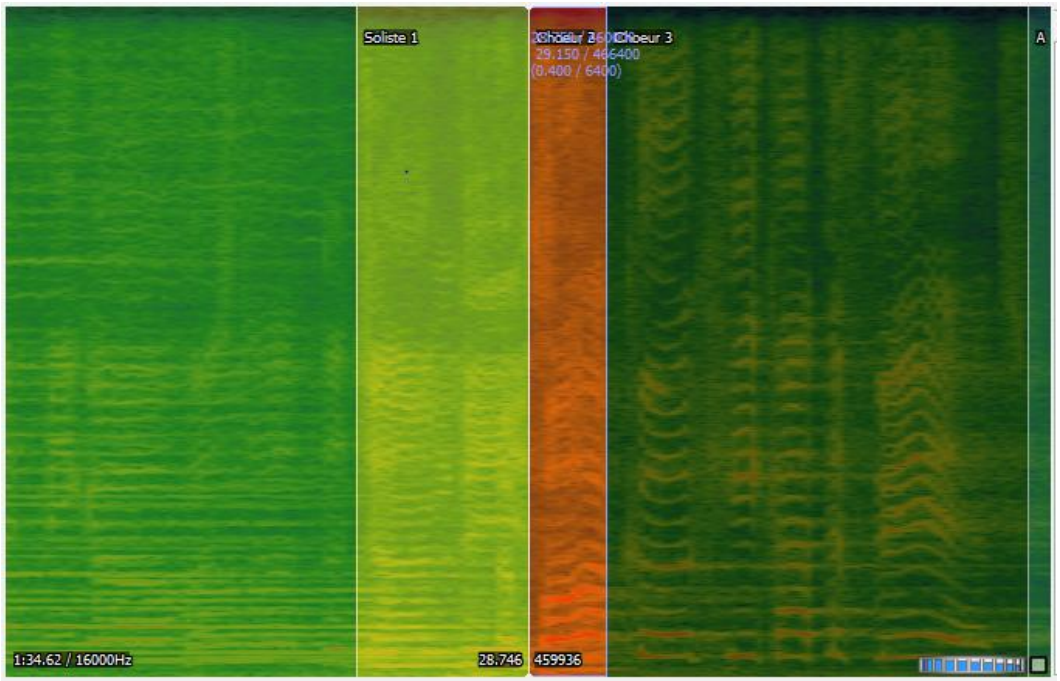

Figure 3. 8 – Illustration d'un cas d'alternance rapide sur un extrait de 5,5 secondes du fichier « sloopJohnB_2_eval ». Le segment en rouge représente un tour de chant de 0,4 seconde.

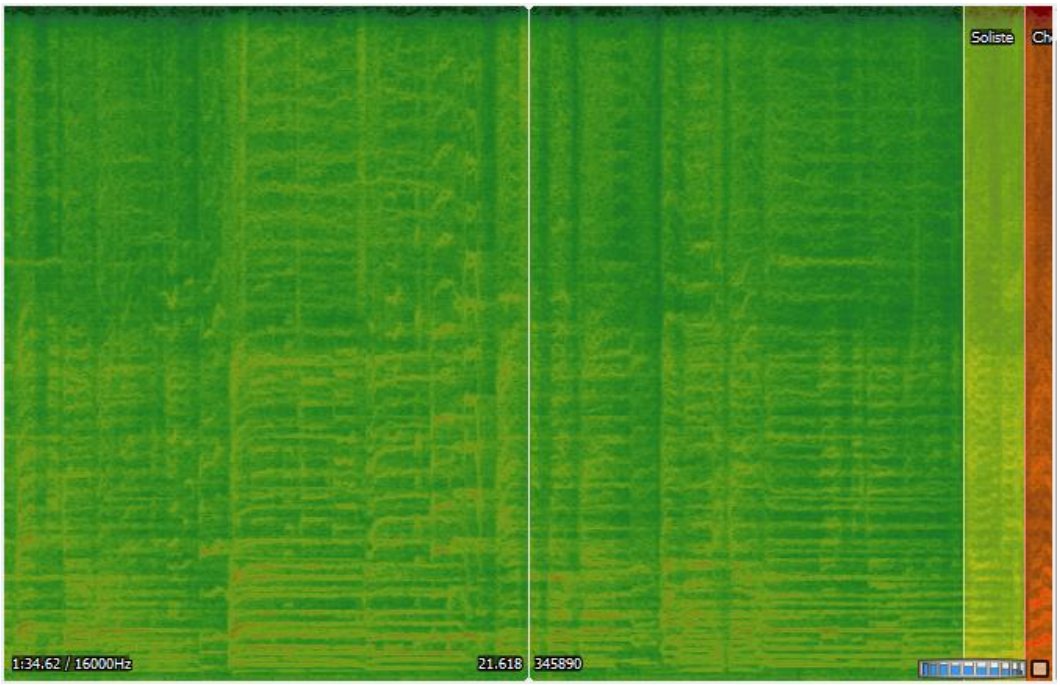

Figure 3. 9 – Illustration d'un cas d'alternances très rapides sur un extrait de 15 secondes du fichier « sloopJohnB_2_eval ». Le premier segment en vert représente une zone constituée de plusieurs alternances très rapides.

3.4.2.6. Regroupement en chanteurs

Après avoir précisé les frontières de début et de fin de chaque tour de chant, nous devons regrouper ces segments. Pour cela, il suffit d'étiqueter tous les segments chantés par un même groupe de chanteur(s) avec le même identifiant (la même étiquette). La Figure 3.10 montre le processus du regroupement.

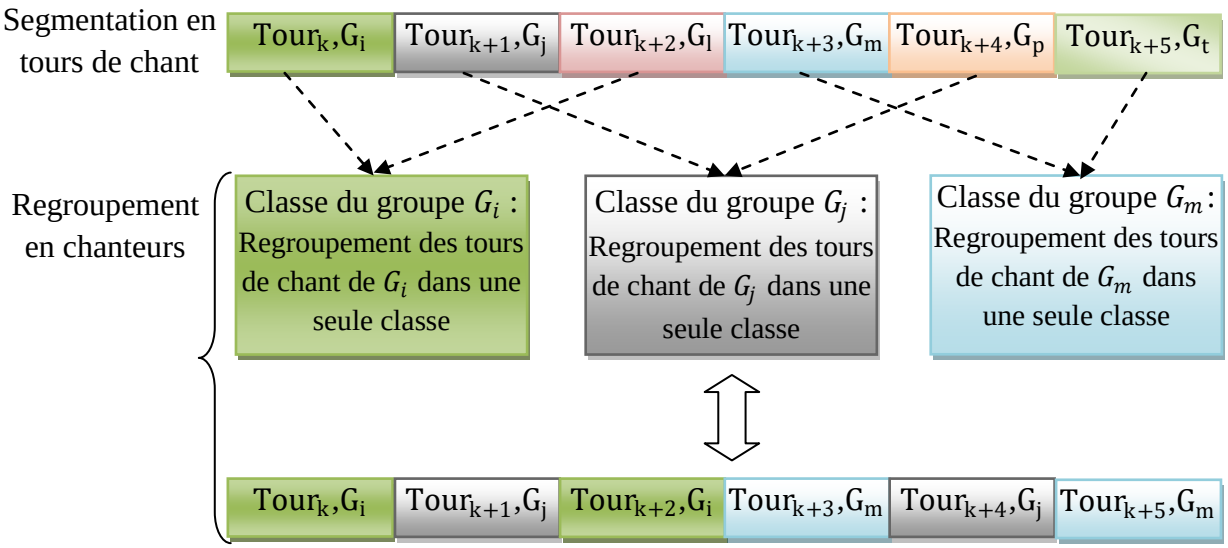

Figure 3. 10 – Illustration du processus du regroupement en chanteurs.

A partir de la sortie de segmentation en tours de chant, nous effectuons le regroupement des tours de chant chantés par le même groupe dans une seule classe en attribuant la même étiquette qui correspond à l'identifiant donné au groupe de chanteur(s) lors de son premier tour de chant dans l'enregistrement.

3.5. Critères d'évaluation

Pour évaluer la performance des systèmes de segmentation en tours de chant et regroupement en chanteurs, nous utilisons les mêmes métriques que celles utilisées pour la segmentation en tours de parole et le regroupement en locuteurs : la précision, le rappel et la F-mesure pour la segmentation et le *Diarization Error Rate* (DER) pour le regroupement. Nous rappelons dans cette section la méthode de calcul de ces critères.

3.5.1. Précision, Rappel et F-mesure

La précision, le rappel et la F-mesure sont des mesures classiques utilisées pour évaluer la segmentation en locuteurs, c'est à dire le positionnement des frontières des tours de parole. Nous utilisons ces critères pour évaluer la segmentation en chanteurs, c'est à dire les frontières des tours de chant. Le rappel permet de calculer le taux de détection des frontières pertinentes parmi toutes les vraies frontières à détecter par le système. La précision est proportionnelle au nombre de fausses frontières (Fausses Alarmes) détectées par le système par rapport au nombre de frontières qui doivent être détectées. La F-mesure est une moyenne harmonique des deux scores de rappel et précision, qui permet d'évaluer la performance globale du système en tenant compte de sa capacité à détecter les frontières pertinentes par rapport à la tâche demandée, et aussi de sa précision. Les expressions de la précision, rappel et F-mesure sont définies de la façon suivante :

$$\text{Précision} = \frac{\sum \text{frontière correctement détectée}}{(\sum \text{frontière correctement détectée} + \sum \text{fausse frontière détectée})}$$

$$\text{Rappel} = \frac{\sum \text{frontière correctement détectée}}{(\sum \text{frontière correctement détectée} + \sum \text{vraie frontière non détectée})}$$

$$\text{F-mesure} = \frac{2 \times \text{Précision} \times \text{Rappel}}{(\text{Précision} + \text{Rappel})}$$

Lors du calcul de ces métriques, une tolérance de « T_{min} » est permise sur les frontières des segments de la référence. Lors de l'évaluation (chapitres suivants), nous donnerons alors la valeur de $T_{min} = 0{,}5\ s$.

3.5.2. Diarization Error Rate

Le *Diarization Error Rate* (DER) représente la métrique usuelle pour évaluer un système complet de regroupement en locuteurs (NIST, 2003). Nous utilisons ce même critère pour

évaluer notre étape de regroupement en chanteurs. Le DER est alors défini par la somme de trois types d'erreurs (Anguera Miro, 2006) :

- $TE_{Substitution}$: pourcentage de temps de chant d'un groupe de chanteurs incorrectement identifié (par exemple G_1 chante alors que le système détecte G_2). Il se calcule comme suit :

$$TE_{Substitution} = \frac{\sum_{s=1}^{S} dur(s) \times \left(\min\left(N_{Réf}(s), N_{Sys}(s)\right) - N_{Correct}(s)\right)}{\sum_{s=1}^{S} dur(s) \times N_{Réf}(s)}$$

- $TE_{Fausse\ Alarme}$: pourcentage de temps de chant détecté par le système alors qu'il n'existe pas. Cette erreur se calcule comme suit :

$$TE_{Fausse\ Alarme} = \frac{\sum_{s=1}^{S} dur(s) \times \left(N_{Sys}(s) - N_{Correct}(s)\right)}{\sum_{s=1}^{S} dur(s) \times N_{Réf}(s)}$$

- $TE_{Manqué}$: pourcentage de temps de chant non détecté par le système alors qu'il est présent dans la référence (par exemple G_1 chante alors que le système ne le détecte pas). Cette erreur se calcule comme suit :

$$TE_{Manqué} = \frac{\sum_{s=1}^{S} dur(s) \times (N_{Réf}(s) - N_{Sys}(s))}{\sum_{s=1}^{S} dur(s) \times N_{Réf}(s)}$$

Avec :

- ✓ $N_{Réf}(s)$ le nombre de groupes de chanteur(s) dans le segment s annotés manuellement (référence),
- ✓ $N_{Sys}(s)$ le nombre de groupes de chanteurs dans le segment s annotés automatiquement (par le système),
- ✓ $N_{Correct}(s)$ le nombre de groupes de chanteurs détectés correctement par le système,
- ✓ $dur(s)$ la durée du segment s,
- ✓ S le nombre total des segments.

Pour calculer le DER, nous utilisons l'outil développé par NIST (National Institute of Standards and Technology)[4]. Nous accordons pour le calcul de ce critère une tolérance T_{min} sur les frontières des segments, la même que précédemment (étape de segmentation).

3.6. Conclusion

Dans ce chapitre, nous avons commencé par définir le terme clé de notre travail qui est « **tour de chant** » que nous avons considéré comme le résultat d'un mécanisme d'alternance entre différents groupes de chanteur(s). Ensuite, nous avons présenté notre contexte de validation.

[4] http://www.nist.gov/speech/tools/index.htm

Puis, nous avons décrit une partie de nos enregistrements (corpus « studio »). Nous avons précisé les différentes conventions d'annotation en tours de chant que nous avons mises en place, en analogie avec la segmentation en tours de parole. Ces conventions consistent à considérer que :

- un nouveau tour de chant est caractérisé d'une frontière de début de segment et de sa durée,
- un tour de non chant est identifié lorsqu'il s'agit d'une pause longue de non-chant. Il est marqué comme un tour de non-chant et caractérisé aussi par une frontière de début et de sa durée,
- une courte pause est incluse avec les tours de chant,
- une courte superposition entre deux groupes de chanteur(s) implique l'existence d'un nouveau tour de chant composé de l'union des deux groupes,
- des alternances très rapides dont la durée est inférieure à un certain seuil sont considérées comme un seul tour de chant.

Ensuite, nous avons illustré le processus d'annotation pour le regroupement en chanteurs. Enfin, nous avons présenté les métriques que nous utilisons pour l'évaluation de notre système qui sont les mêmes que celles développées pour la parole : Précision, Rappel et F-mesure pour la segmentation et le DER pour le regroupement.

Chapitre 4

Segmentation en tours de chant

Sommaire

4.1. Introduction

Pour des tâches d'indexation de musique, il peut s'avérer intéressant de détecter les tours de chant. Il est rappelé qu'un tour de chant est défini, pour nous, comme une alternance entre deux groupes de chanteurs, composés chacun d'un seul chanteur ou de plusieurs chanteurs. La localisation des tours de chant peut être utilisée dans plusieurs applications, par exemple pour la recherche des segments chantés par un groupe de chanteur dans un enregistrement musical donné ou même dans de grandes archives de la musique. En effet, parcourir tout un document audio, qui peut contenir des données commerciales ou des segments de parole, est une perte de temps et n'est pas pratique pour l'utilisateur qui cherche des informations pertinentes. Dans ce cas, la détection des tours de chant peut être une première étape afin d'« estimer le nombre de chanteurs présents dans un document sonore » ou « reconnaître qui chante ».

Dans la littérature, de nombreux travaux existent sur le chant mais très peu, à notre connaissance, sur la segmentation en tours de chant. Néanmoins, en parole, il existe beaucoup d'études sur la segmentation en tours de parole dont la problématique de détection d'alternance entre groupe de locuteurs semble assez proche de notre travail. Le lecteur pourra se référer au Chapitre 2 pour une description de l'état de l'art sur le chant ainsi que sur la segmentation en tours de parole.

En partant de l'idée que la parole et le chant partagent certaines caractéristiques et en s'inspirant des travaux en segmentation en tours de parole, nous avons développé un système de segmentation en tours de chant qui permet de détecter les points de changement entre chanteurs en utilisant des modèles probabilistes.

Nous avons exploité les travaux effectués en segmentation en tours de parole pour réaliser la même tâche pour le chant, en s'appuyant sur le Critère d'Information Bayésien (BIC). Le BIC comme son nom l'indique, se place dans un contexte bayésien de sélection de modèles. Variante du critère d'Akaïke, il est utilisé dans de nombreux contextes applicatifs et ce depuis très longtemps (Akaike, 1974), (Schwarz, 1978). Plus récemment, il est au cœur de nombreux travaux de segmentation sonore (Cettolo, et al., 2005), (Chen, et al., 1998a), (Delacourt, et al., 2000) et bien des systèmes de segmentation et de regroupement en locuteurs, actuels et performants, se fondent sur le BIC. Le calcul de ce critère pour la phase de segmentation et la phase de regroupement est détaillé dans le deuxième chapitre. J'ai travaillé sur l'adaptation de ce critère au contexte du chant et aussi sur le choix de paramètres acoustiques pour la paramétrisation du chant.

Ce chapitre est divisé en cinq parties. Nous présentons, dans une première section, les limites des méthodes disponibles en segmentation en locuteurs lorsque nous les appliquons sur le chant. Une deuxième section est dédiée à la présentation de l'algorithme de référence de segmentation par BIC. Une description globale de notre système de segmentation en tours de chant ainsi que nos contributions font l'objet de la troisième section. Dans la quatrième section, nous décrivons les paramètres acoustiques et les hyper-paramètres de l'algorithme.

Les résultats de notre système de segmentation en tours de chant sont exposés dans la cinquième section.

4.2. Limites des méthodes de segmentations « statiques »

Comme nous essayons de réaliser un système de segmentation en tours de chant par analogie aux systèmes de segmentation en tours de parole, nous avons appliqué dans un premier temps, des méthodes disponibles pour la segmentation en tours de parole sur des enregistrements de chant. Ces méthodes sont celle de l'IRIT qui est accessible vu notre lieu du travail et celle du LIUM qui est disponible en ligne[5]. Elles sont toutes les deux décrites dans le chapitre d'état de l'art.

Ces deux approches utilisent le BIC pour le découpage du signal en des segments acoustiquement homogènes par locuteur. Ce critère nécessite de fixer les valeurs de deux paramètres : la taille de la fenêtre d'analyse N et le facteur de pénalité λ. Les deux méthodes de segmentation du LIUM et de l'IRIT, à l'instar d'autres approches qui utilisent le BIC, ajustent les paramètres N et λ sur un corpus de développement. En appliquant le système IRIT (El-Khoury, et al., 2009), utilisant une fenêtre d'analyse égale à 2 s et en faisant varier la valeur du facteur de pénalité λ, la meilleure performance atteignit 33% de F-mesure (cf. Tableau 4. 1) sur le sous-ensemble de développement (DEV) du corpus « studio ». Une performance de 38% (cf. Tableau 4. 1) a été obtenue en appliquant le système LIUM (Meignier, et al., 2009) sur le même corpus. Ces performances sont faibles par rapport aux scores obtenus en segmentation en locuteurs par ses systèmes (F-mesure supérieure à 80%). Cela est principalement dû au fait que la recherche d'une taille optimale de la fenêtre d'analyse, indépendante de l'enregistrement considéré, s'est révélée difficile. En effet, contrairement à la parole pour laquelle la taille de la fenêtre d'analyse est constante, pour le chant, elle varie énormément d'un enregistrement à l'autre. Pour les enregistrements contenant des courtes alternances dont la durée est inférieure à 0,4 s, une taille de fenêtre petite est plus adaptée pour ne pas avoir un problème de sous-segmentation (manque de détection de frontières). Par contre, une taille de fenêtre petite n'est pas adaptée aux enregistrements contenant des longues alternances, dont la durée est supérieure à 10 s, car cela engendre un problème de sur-segmentation.

Comme fixé le paramètre N est quasiment impossible dans un contexte de chant, nous avons cherché une autre version de segmentation par BIC qui nous évite de déterminer *a priori* la taille de la fenêtre d'analyse et qui est détaillée dans la section suivante (4.3).

A ce problème de variabilité de la taille de la fenêtre d'analyse, se rajoute le problème de variabilité du facteur de pénalité λ. Le Tableau 4. 1 montre les résultats des deux systèmes IRIT et LIUM pour deux exemples du DEV et pour la totalité du DEV du corpus « studio » en utilisant une taille de fenêtre fixe de 2 s et le meilleur λ possible, ajusté en faisant varier sa valeur dans l'intervalle $[2 \quad 10]$ avec un pas de 0.05.

[5] http://www-lium.univ-lemans.fr/diarization/doku.php/welcome

La valeur du facteur de pénalité est assez différente d'un fichier à l'autre et aussi d'un système à l'autre. Pour le fichier « 03-Mayingo_dev », les deux systèmes ne détectent pas les changements entre les chœurs et ils considèrent qu'il s'agit d'un même groupe de chanteur pour les différents chœurs. Donc, un problème de sous-segmentation se pose, ce qui justifie le taux de rappel faible de 57,1% (respectivement 50%) avec le système IRIT (respectivement LIUM). Pour le fichier « Arranoak_Bortietan_dev », deux problèmes se posent : un problème de sur-segmentation et un problème de sous-segmentation. Le problème de sur-segmentation réside dans le fait que les deux systèmes ont tendance à segmenter les changements de note au sein d'un même tour de chant. Cela induit la présence de plusieurs fausses alarmes, ce qui explique la faible précision obtenue avec les deux systèmes (une précision de 37,5% avec le système IRIT et de 30% avec le système LIUM). Le problème de sous-segmentation est engendré à cause de la présence de plusieurs transitions chant-silence dans ce fichier. Dans cette situation, les deux systèmes ne détectent pas les frontières de fin des tours de chant qui correspondent au début du silence, et donc ils incluent le silence avec les tours de chant. Cela implique l'absence de détection de plusieurs frontières, ce qui induit à un taux de rappel faible de 25% avec le système IRIT et 23,1% avec le système LIUM.

Nous notons aussi que la F-mesure est très variable d'un enregistrement à l'autre avec les deux systèmes, ce qui explique les scores obtenus : une F-mesure de 61,5% (respectivement 66,6%) avec le système IRIT (respectivement LIUM) pour le fichier « 03-Mayingo_dev », et une F-mesure de 30% (respectivement 26,1%) avec le système IRIT (respectivement LIUM) pour le fichier « Arranoak_Bortietan _dev ».

Tableau 4. 1 – Résultats des systèmes IRIT et LIUM de segmentation en tours de parole sur deux exemples du DEV du corpus « studio » et sur la totalité du DEV.

Système	**Fichier « 03-Mayingo_dev »**		**Fichier «Arranoak_Bortietan_dev»**		**DEV**	
	IRIT	**LIUM**	**IRIT**	**LIUM**	**IRIT**	**LIUM**
λ	2,00	5,45	2,20	2,00	Best λ	
Précision (%)	66,7	100	37,5	30	42,2	51,5
Rappel (%)	57,1	50	25	23,1	27,3	30,3
F-mesure (%)	61,5	66,6	30	26,1	33,1	38,1

4.3. Segmentation « dynamique »

Comme les versions qui utilisent une taille de fenêtre fixe semblaient peu performantes, limitées à un contexte de parole, nous avons décidé d'utiliser une autre version de segmentation par BIC dans laquelle la taille de la fenêtre d'analyse est dynamique, tels que les travaux décrits dans (Delacourt, et al., 1999) et (Cettolo, et al., 2005). En effet, dans ce contexte, la fenêtre d'analyse augmente tant qu'aucune frontière potentielle n'est trouvée. Ce

procédé s'inspire d'études en segmentation de la parole, et son déroulement général est illustré dans la Figure 4. 1.

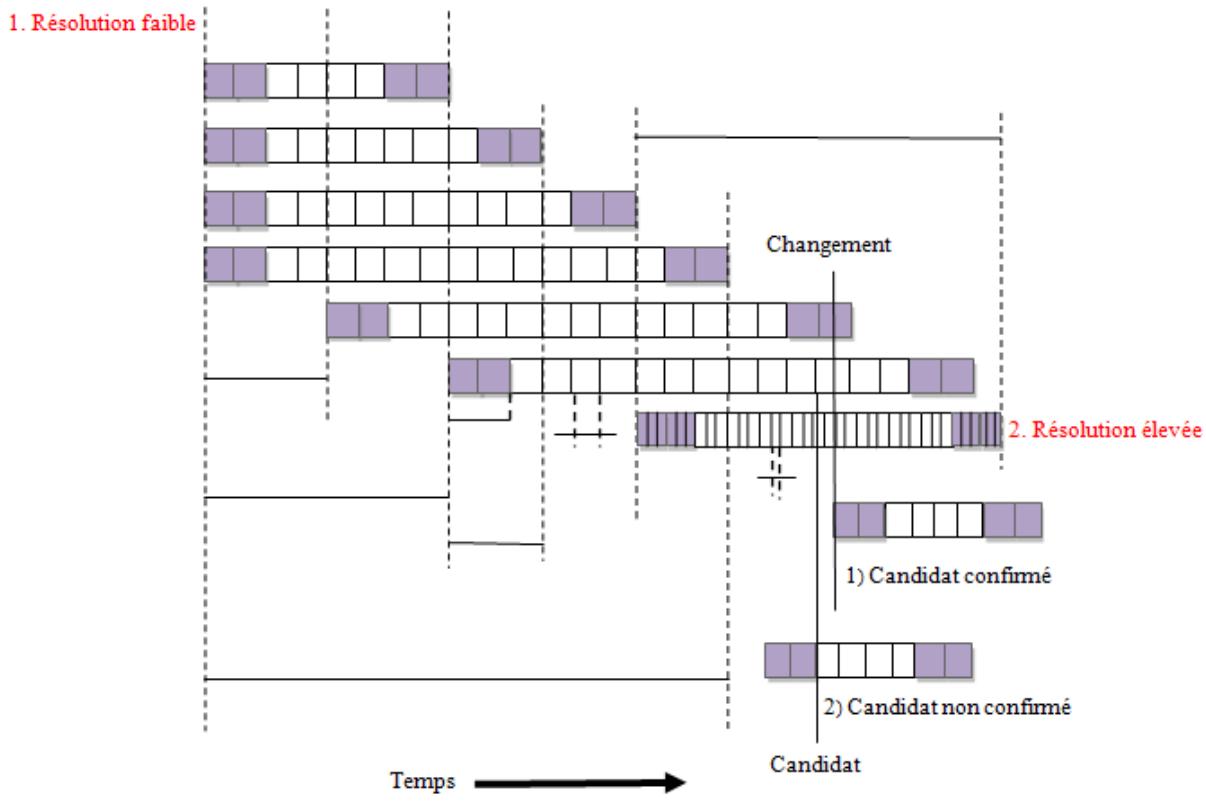

Figure 4. 1 – Illustration générale de l'algorithme de segmentation par BIC de (Cettolo, et al., 2005).

La recherche se déroule en deux temps, impliquant deux résolutions temporelles différentes :

1. Résolution « faible » : le déroulement de cette résolution est détaillé dans la Figure 4. 2. La longueur initiale de la fenêtre d'analyse est fixée à N_{min}. Elle augmente de $\Delta N_{augmentation}$ tant que le test d'hypothèse ΔBIC ne valide aucune frontière interne. Les valeurs de ΔBIC sont calculées à intervalles réguliers pour des valeurs échantillonnées de t, à savoir une fois toutes les δ_1 observations. Si aucune frontière n'est détectée lorsque une valeur N_{max} est atteinte, la fenêtre d'analyse est décalée de $\Delta N_{décalage}$ et l'analyse est réinitialisée.

2. Résolution « élevée » : si une frontière potentielle est détectée, une fenêtre de longueur $N_{seconde}$ est centrée sur cette frontière et les valeurs de ΔBIC sont calculées au sein de cette fenêtre à une résolution haute, toutes les δ_h observations, afin d'affiner la position de cette frontière. le déroulement de cette résolution est détaillé dans la Figure 4. 3.

Nous imposons que toute frontière ne produise aucun segment d'une durée inférieure à N_{marge}, ce qui implique qu'aucune frontière n'est recherchée entre les zones $[1, N_{marge}]$ et $[N - N_{marge}, N]$.

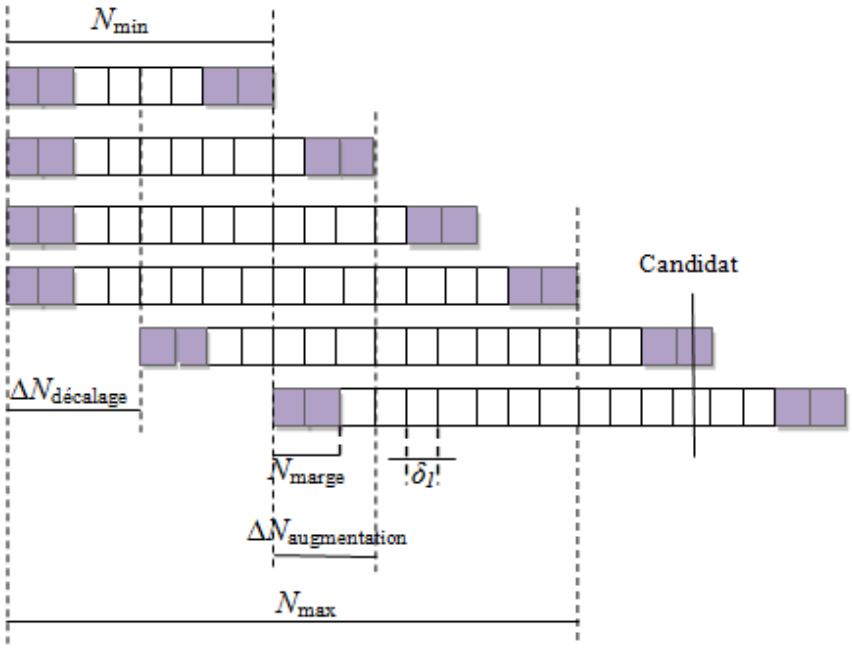

Figure 4. 2 – Illustration de la résolution « faible » de l'algorithme de segmentation par BIC de (Cettolo, et al., 2005).

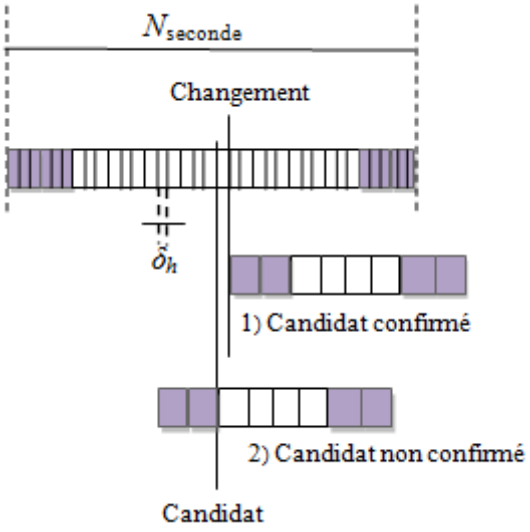

Figure 4. 3 – Illustration de la résolution « élevée » de l'algorithme de segmentation par BIC de (Cettolo, et al., 2005).

4.4. Présentation de notre méthode de segmentation

Nous présentons ici notre méthode de segmentation en tours de chant. Son architecture est illustrée dans la Figure 4. 4. Des contributions ont été apportées sur chacune des 3 points :

- Choix des paramètres acoustiques : la paramétrisation consiste à extraire des caractéristiques du signal. Pour ce point, nous avons testé les MFCC, les chromas et les FBANK.

- Choix des hyper-paramètres du BIC : dans cette étape, nous avons ajusté les valeurs des différents paramètres de l'algorithme de segmentation par BIC de (Cettolo, et al., 2005) au contexte du chant.

- Décision Consolidée *A Posteriori* (DCAP) : cette méthode a été proposée pour pallier au problème de variabilité du facteur de pénalité et pour éviter le choix *a priori* de sa valeur.

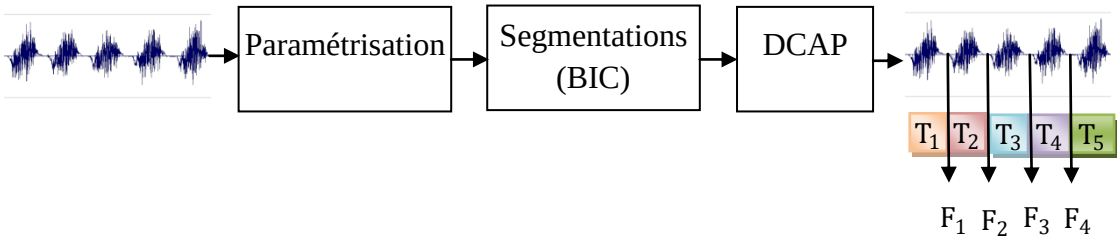

Figure 4. 4 – Architecture de notre système de segmentation en tours de chant.

La paramétrisation est présentée dans la section 4.5.4 dans laquelle nous étudions l'influence et le choix des paramètres acoustiques. Le choix et l'adaptation des hyper-paramètres de l'algorithme de segmentation par BIC sont détaillés dans la section 4.4.1. La méthode DCAP est présentée dans la section 4.4.2.

4.4.1. Adaptation de l'algorithme de référence pour la segmentation en tours de chant

Le système de segmentation par BIC, utilisant une taille de fenêtre d'analyse « dynamique », a montré une amélioration importante dans les résultats de segmentation par rapport à ceux obtenus avec les systèmes utilisant une taille de fenêtre fixe (systèmes IRIT et LIUM). Contrairement aux systèmes IRIT et LIUM qui ne possèdent que deux paramètres (N et λ) à déterminer, cet algorithme contient plusieurs paramètres qui doivent être adaptés et ajustés par rapport à notre corpus de chant. Ainsi, nous avons utilisé un corpus de développement (DEV), détaillé dans le troisième chapitre, pour déterminer :

- N_{min} : la taille minimale de la fenêtre de recherche d'une frontière est fixée à 0,8 secondes tandis que sa longueur maximale N_{max} est de 5 secondes,
- $\Delta N_{augmentation}$: le nombre d'observations ajoutées à la fenêtre de détection tant qu'il n'y a pas de frontière de segments détectée et tant que la taille maximale n'est pas atteinte, correspond à 0,5 secondes,
- $\Delta N_{décala\,ge}$: le décalage de la fenêtre sur le signal lorsque la taille maximale est atteinte et qu'aucune frontière potentielle n'est découverte correspond à 0,4 secondes,
- N_{marge} : la taille minimale d'un segment de chant est de 0,7 secondes,
- $N_{secon\,de}$: la taille de la fenêtre d'analyse fine correspond à 1,2 secondes.

L'étape de faible résolution temporelle du calcul du ΔBIC utilise 1 trame sur 5 : $\delta_1 = 5$ (soit 50 millisecondes) tandis que la haute résolution considère toutes les trames : $\delta_h = 1$ (soit 10 millisecondes de précision).

En appliquant cet algorithme avec ces valeurs de paramètres sur des enregistrements de chant et en utilisant les paramètres MFCC comme pour les systèmes « statiques » précédents (IRIT et LIUM), la meilleure performance trouvée sur le DEV est de 59% de F-mesure. L'augmentation des performances est d'environ 26% et 21% (cf. Tableau 4. 1), respectivement par rapport aux systèmes IRIT et LIUM.

4.4.2. DCAP : Décision Consolidée *A Posteriori*

Après avoir résolu le problème de variabilité de la taille de la fenêtre d'analyse N en implémentant une méthode de segmentation utilisant une taille dynamique, nous avons essayé de déterminer la valeur du facteur de pénalité λ sur le corpus de développement (DEV). Mais, le choix d'une valeur optimale unique pour tous les enregistrements s'est avéré difficile car la valeur de λ varie grandement en passant d'un fichier à un autre.

Afin d'éviter ce problème de variabilité et le choix *a priori* du facteur de pénalité, nous avons proposé une méthode basée sur la fusion des sorties de plusieurs segmentations, qui est inspirée des travaux utilisés pour la parole (Abad, et al., 2013). En effet, les auteurs utilisent une technique de fusion pour améliorer la performance de leur système de détection des mots clés. Cette technique est basée sur un vote majoritaire : les sorties de plusieurs systèmes hétérogènes de détection sont mélangées et un ensemble de candidats est définit en ne conservant que les segments détectés par la majorité des systèmes. Une combinaison des scores est réalisée pour produire un seul score par détection de candidat en prenant la moyenne des scores de tous les systèmes. Cette technique de fusion a amélioré la performance du système de détection des mots clés d'environ 13%.

Nous avons proposé une méthode simple de fusion, inspirée de cette technique de vote majoritaire, pour améliorer la performance de notre système de segmentation. Nous avons appelé notre méthode : Décision Consolidée *A Posteriori* (DCAP). Celle-ci est illustrée sur la Figure 4. 5.

Tout d'abord, nous effectuons plusieurs segmentations (M segmentations) en faisant varier la valeur du coefficient de pénalité λ. Ensuite, toutes les frontières obtenues à partir de ces segmentations sont fusionnées. Chaque nouveau segment obtenu, dont la durée est inférieure à un certain seuil (qui correspond à la tolérance utilisée lors de l'évaluation de notre système de segmentation) est remplacé par une frontière située à son milieu. Pour décider quelles sont les frontières retenues, un vote est effectué sur les candidats obtenus : une frontière est validée si elle est trouvée par au moins S_0 segmentations parmi les M segmentations obtenues. La valeur de S_0 comprise entre 1 et M est déterminée en utilisant les enregistrements du corpus de développement.

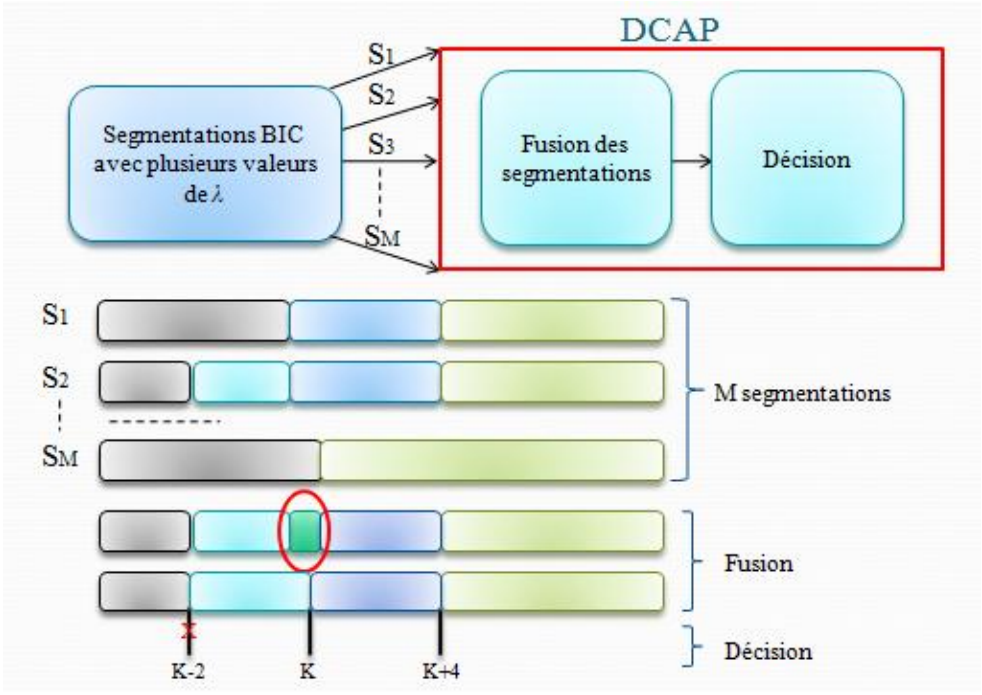

Figure 4. 5 – Illustration de la méthode de Décision Consolidée *A Posteriori*.

4.5. Influence des paramètres acoustiques et hyper-paramètres de l'algorithme

L'étude des paramètres et les expériences présentées dans ce chapitre sont effectuées sur le corpus « studio », i.e. avec des conditions acoustiques contrôlées, détaillé dans le troisième chapitre. Après avoir présenté le système de base, nous analysons l'influence du coefficient de pénalité λ. Puis, nous illustrons l'ajustement des paramètres de la méthode DCAP. Ensuite, une étude comparative de plusieurs paramètres acoustiques est effectuée et une présentation de la nouvelle méthode de paramétrisation est réalisée. Enfin, l'influence du corpus de développement sur l'ajustement du paramètre de la méthode DCAP est montrée.

4.5.1. Evaluation de la première version de segmentation en tours de chant

Comme les systèmes IRIT et LIUM de segmentation en tours de parole utilisent une taille de fenêtre d'analyse constante et que l'utilisation d'une taille de fenêtre fixe n'est pas adaptée à notre contexte de chant, nous avons essayé d'utiliser le tempo (Le Coz, et al., 2012) ainsi que plusieurs combinaisons de sa valeur pour trouver la valeur optimale de N. Malheureusement, ces essais n'ont pas été fructueux et nous avons décidé d'implémenter une première version pour la segmentation en utilisant le BIC avec une taille de fenêtre d'analyse dynamique, détaillé dans la partie 4.3.

Les performances de notre algorithme de segmentation ont été évaluées en termes de Rappel, Précision et F-mesure pondérés sur les durées des fichiers. J'ai utilisé une tolérance de + ou – 50 millisecondes sur les positions des frontières. Par analogie avec les tours de

parole, j'ai commencé par tester les MFCC, qui sont couramment utilisés dans la segmentation en tours de parole et aussi dans certains travaux de traitement de la musique, pour détecter les tours de chant. Les meilleurs résultats ont été obtenus en limitant le vecteur d'observation à 12 coefficients MFCC, sans l'énergie ni le coefficient C_0.

Pour le modèle probabiliste du BIC, nous avons commencé par utiliser des matrices de covariance pleines mais, nous avons remarqué qu'avec cette configuration, nous ne pouvions pas beaucoup faire varier la valeur du facteur de pénalité et ainsi tester des valeurs élevées. Ainsi, notre algorithme avait tendance à sur-segmenter et donc générer beaucoup de fausses alarmes. Par conséquent, nous avons décidé d'utiliser des matrices de covariance diagonales pour le modèle probabiliste du BIC. Ainsi nous pouvons tester des valeurs élevées du facteur de pénalité et contrôler le nombre de fausses alarmes.

Afin de définir une valeur « optimale » du coefficient de pénalitéλ, les performances du système ont été calculées en le faisant varier sur les enregistrements de l'ensemble de développement (DEV). Une valeur standard de λ égale à 1 est utilisée. Avec cette configuration, une F-mesure de 41,7% a été obtenue pour le corpus DEV. Cette performance est meilleure que les segmentations statiques (IRIT et LIUM, cf. section 4.2) mais reste quand même faible. Une analyse plus fine des résultats a montré qu'il y avait beaucoup de fausses alarmes dues à une sur-segmentation selon les notes.

4.5.2. Influence du coefficient de pénalité λ

L'ajustement du coefficient de pénalité s'est avéré délicat ; nous illustrons et analysons dans cette partie sa pertinence ainsi que sa grande sensibilité aux variations de contenus des enregistrements.

Le rôle du coefficient de pénalité λ dans le critère BIC est de pénaliser une modélisation trop complexe : dans le cadre gaussien et multi modèles qui est le nôtre, plus la valeur de λ augmente, plus l'hypothèse H_1 est pénalisée et plus l'insertion d'une frontière est difficile : l'algorithme a tendance à moins segmenter. Globalement, choisir la bonne valeur de λ revient à trouver le bon compromis entre Rappel et Précision. Nous avons remarqué que la performance varie de manière importante en fonction de ce facteur.

La valeur de λ varie d'un enregistrement à l'autre dès lors que nous cherchons à l'optimiser sur un enregistrement donné. Pour certains enregistrements, nous trouvons de bonnes performances avec des valeurs proches de 10 (lorsque les segments attendus sont longs). Les valeurs moins élevées, de l'ordre de 2, se révèlent meilleures là où les segments attendus sont plus courts. Cela engendre une variabilité importante de la performance globale de notre système en fonction de λ.

A des fins de comparaison, nous avons, pour chaque enregistrement du corpus DEV et du corpus EVAL, déterminé la meilleure valeur de F-mesure obtenue en faisant varier le coefficient de pénalité : nous appellerons ce système artificiel, le système « *oracle* ». Les

performances globales du système « *oracle* » sur le DEV sont données dans le Tableau 4. 2. Sur le DEV, la F-mesure du système « *oracle* » atteint 59,5% (elle n'était que de 41,7% avec une valeur standard de λ égale à 1). Nous trouvons un écart de performance d'environ 30% relatifs (18% absolus). Cette différence confirme la nécessité de ne pas fixer *a priori* le coefficient de pénalité, quitte à inclure un post-traitement pour obtenir une segmentation *a posteriori*.

4.5.3. Ajustement des paramètres de la méthode DCAP

Comme la valeur du coefficient de pénalité varie beaucoup en passant d'un enregistrement de chant à un autre, nous avons effectué plusieurs expériences en essayant des valeurs de λ comprises entre $0{,}05$ et 10. Nous avons remarqué que les valeurs de λ qui donnent de meilleures performances sur le DEV se situent entre 2 et 10. Ce qui nous a conduit à réduire l'intervalle de la variation de λ de $[0{,}05 \quad 10]$ à $[2 \quad 10]$.

Pour éviter le choix *a priori* du facteur de pénalité, nous appliquons notre méthode DCAP en réalisant plusieurs segmentations obtenues à partir de la variation de λ sur l'intervalle$[2 \quad 10]$ avec un pas de $0{,}05$. Ainsi, nous obtenons 161 segmentations d'un même enregistrement. La valeur de S_0 comprise entre 1 et 161 est déterminée sur le corpus de développement et elle est égale au nombre de systèmes qui permet de donner la meilleure performance en termes de F-mesure, en tolérant un écart de $0{,}5$ secondes. Nous présentons dans la section suivante les valeurs trouvées de S_0 pour les différents paramètres acoustiques testés.

4.5.4. Influences des paramètres acoustiques

L'application de notre méthode de segmentation sur des enregistrements musicaux de type « studio »en utilisant les MFCC, n'est pas totalement satisfaisante. En effet, les performances sont moyennes : des F-mesures de 59,5% pour le système « *oracle* » (la performance maximale que notre système peut atteindre) et de 57,0% pour notre système DCAP (cf. Tableau 4. 2 et Tableau 4. 4 pour l'évaluation sur le DEV). Ces résultats nous ont conduit à remettre en question les paramètres acoustiques utilisés.

Afin d'améliorer les performances de notre système, nous avons effectué des expériences avec d'autres paramètres acoustiques que les MFCC. Ces expériences nous ont permis de trouver de meilleurs paramètres ainsi que la méthode de paramétrisation adaptée à notre tâche. L'architecture finale de notre système de segmentation est illustrée sur la Figure 4. 6.

Dans cette partie, nous illustrons les performances avec différents types de paramètres et nous présentons une stratégie de paramétrisation. Elle consiste à sélectionner automatiquement les bandes fréquentielles (coefficients FBANK) qui contiennent le plus d'informations par rapport à l'enregistrement traité.

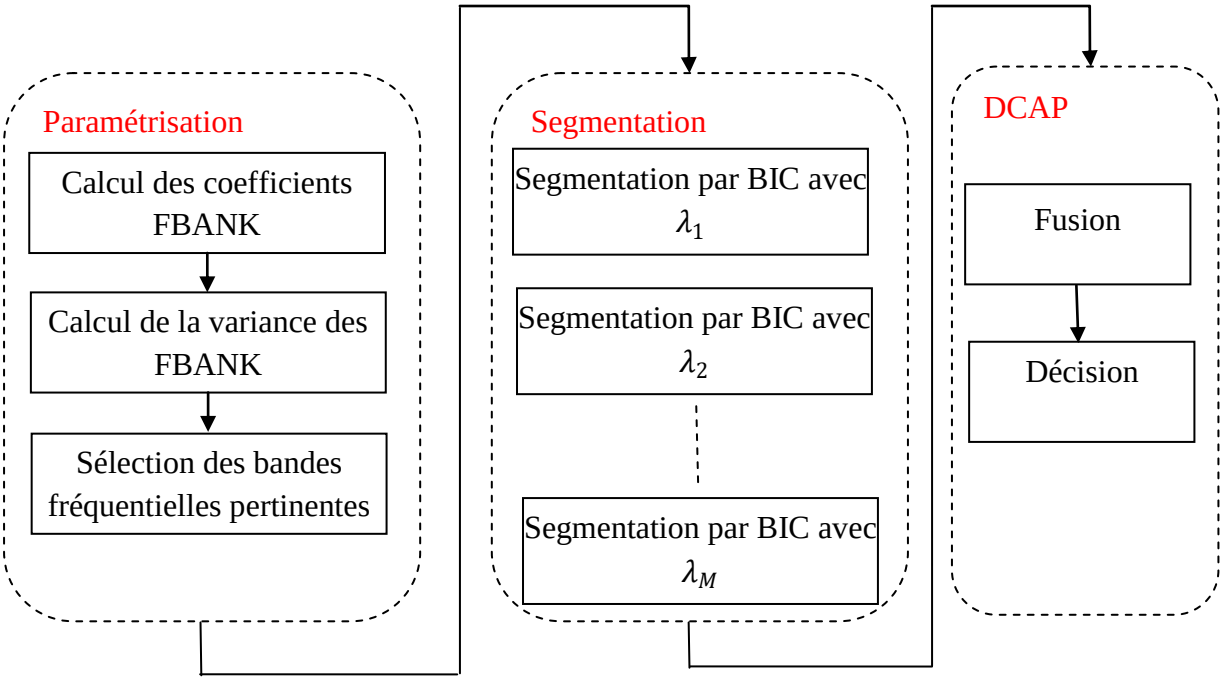

Figure 4. 6 – Architecture finale du système complet de segmentation en tours de chant.

4.5.4.1. Etude de différents paramètres acoustiques

Par analogie avec les tours de parole, nous avons testé les paramètres acoustiques couramment utilisés en traitement automatique de la parole. Ainsi, nous avons appliqué notre système de segmentation en tours de chant sur plusieurs configurations de descripteurs : MFCC, accompagnés ou non de l'énergie et des dérivées premières et secondes, PLP (Perceptual Linear Prediction), RASTA-PLP et FBANK. Nous avons aussi essayé des paramètres musicaux tels les chromas. Nous ne présentons ici que les meilleurs systèmes, utilisant les paramètres suivants :

- 12 MFCC,
- 12 chromas,
- un nombre variable de FBANK.

Le Tableau 4. 2 illustre les résultats du système « *oracle* » avec les MFCC, les chromas et les FBANK sur le corpus « studio ». Ainsi, en utilisant12 MFCC, 12 chroma et 24 FBANK, les valeurs de F-mesure sont respectivement de 59,5%, 66,8% et 78,7%. En utilisant 24 FBANK, la précision augmente de 1,9 et 1,5 fois plus qu'avec 12 MFCC et 12 chromas, respectivement. En effet, le nombre de fausses alarmes diminuent fortement pour tous les fichiers du corpus. En utilisant 12 MFCC, le système produit beaucoup de fausses alarmes : détections de longues notes tenues. Le nombre de fausses alarmes élevé avec les chromas est dû au fait que ces paramètres suivent principalement la mélodie.

Tableau 4. 2 – Performances du système « oracle » de segmentation en tours de chant sur le DEV du corpus « studio ».

Paramètres	Précision (%)	Rappel (%)	F-mesure (%)
12 MFCC – « *oracle* »	47,3	80	59,5
12 Chromas – « *oracle* »	59,4	76,3	66,8
24 FBANK – « *oracle* »	88,1	71,1	78,7
12 FBANK – « *oracle* »	91,5	84	87,5
FBANK sélectionnés – « *oracle* »	82,6	94,3	88,1

4.5.4.2. Choix des coefficients FBANK

Les expériences successives nous ont amenés à rechercher quelles bandes de fréquence étaient les plus pertinentes dans cette étude du chant et par voie de conséquence à remettre en cause la largeur des bandes fréquentielles.

- **Sélection de la bande fréquentielle supérieure**

Comme les FBANK ont donné les meilleures performances, nous avons fait plusieurs expériences pour découvrir les caractéristiques de ces bandes fréquentielles. Nous avons testé plusieurs configurations en faisant varier le nombre des coefficients FBANK et la valeur du facteur de pénalité afin d'avoir la meilleure valeur de λ pour chaque enregistrement avec chacune des configurations. Ainsi, nous avons obtenu la courbe illustrée dans la Figure 4. 7 qui représente la variation de la meilleure performance (système « *oracle* ») obtenue sur le DEV, en fonction du nombre de bandes fréquentielles (FBANK) utilisées à l'entrée de notre système de segmentation.

En observant le comportement de cette courbe, nous avons remarqué qu'utiliser des bandes de fréquence au-delà des 11 ou 12 premières dégrade la performance. Pour cette raison, nous avons décidé d'utiliser seulement les 12 premiers coefficients FBANK. Ceci nous a permis d'avoir une F-mesure de 87,5% sur DEV. En utilisant 12 FBANK, des gains absolus de 8,8%, 20,7% et 28% ont ainsi été obtenus par rapport aux résultats trouvés respectivement avec 24 FBANK, les 12 chromas et les 12 MFCC.

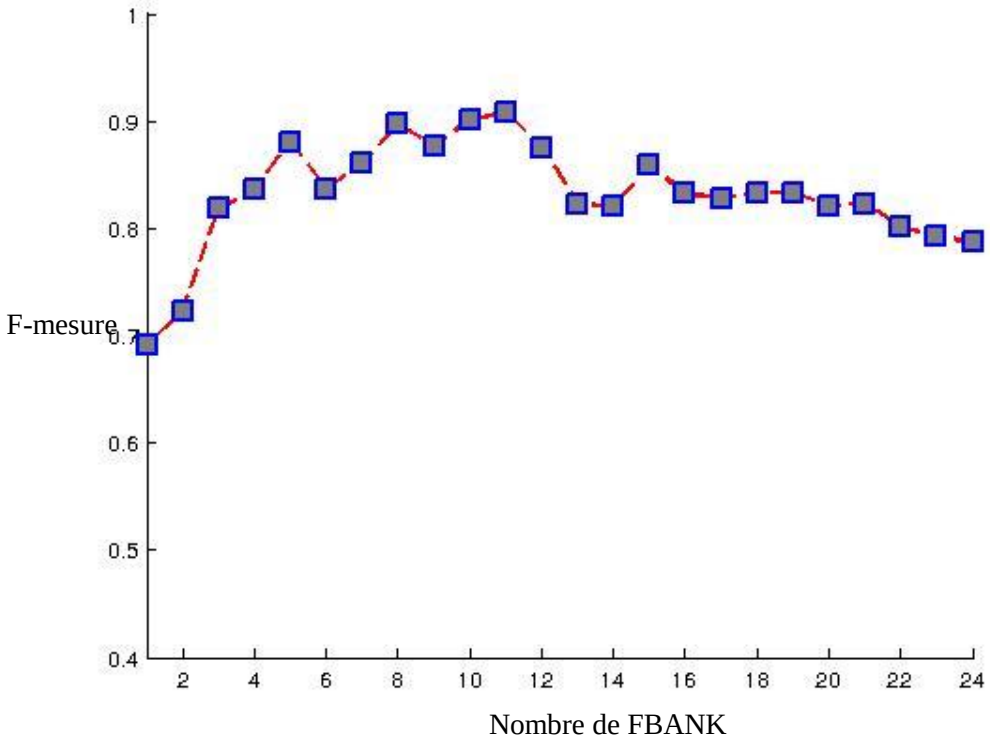

Figure 4. 7 – Variation de la performance du système « oracle » de segmentation en tours de chant en fonction du nombre de bandes (FBANK) sur le corpus DEV.

- **Sélection basée sur la variance**

L'important gain obtenu en utilisant les coefficients FBANK nous a conduit à étudier leurs caractéristiques et chercher une stratégie qui permette de trouver les bandes les plus informatives. Pour cela, nous avons essayé plusieurs techniques. D'abord, nous avons commencé à utiliser les zones de fréquences proches de la valeur du pitch afin de trouver les zones contenant le plus d'informations et ainsi limiter encore le nombre de bandes utilisées. En fonction du fichier à traiter, nous pourrions nous arrêter à la bande correspondante à la valeur du pitch. Malheureusement cette technique n'a pas amélioré les résultats à cause notamment de la présence de sons polyphoniques (par exemple lorsqu'il s'agit d'un chœur) pour lesquelles l'estimation de la fréquence fondamentale n'était pas très fiable. Il se peut aussi que les informations dans les harmoniques soient importantes pour effectuer une segmentation en tours de chant.

Ensuite, nous avons examiné la dispersion des coefficients FBANK et essayé d'extraire les bandes les plus pertinentes en fonction de leur variance. Pour cela, nous avons appliqué une méthode qui permet de ne garder que les coefficients dont la variance dépasse un certain seuil. Or, l'ajustement d'une valeur fixe de ce seuil pour tous les fichiers du DEV s'est avéré difficile. En traçant l'histogramme des variances des bandes, nous avons remarqué que certains coefficients FBANK possèdent des valeurs de variance beaucoup plus importantes que d'autres. En ne sélectionnant que les deux coefficients qui possèdent les variances les plus élevées, nous avons obtenu des performances de l'ordre de 60% : ceci nous a confirmé

l'existence de bandes plus informatives que d'autres et que les bandes possédant des variances importantes étaient pertinentes. En analysant ces histogrammes, nous avons remarqué deux types de comportement : les exemples de la Figure 4. 8 et de la Figure 4. 9 en rendent compte sur deux enregistrements différents du DEV. Nous avons exploité cette observation pour proposer une méthode de paramétrisation qui permet de ne retenir que les coefficients FBANK possédant les plus fortes variances, sous deux conditions.

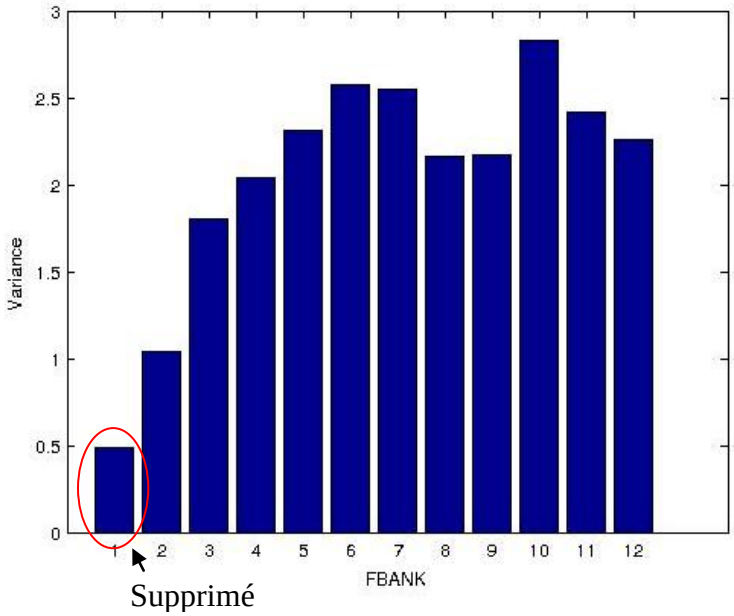

Figure 4. 8 – Variance des coefficients FBANK pour un exemple de 38 s de l'ensemble DEV du corpus « studio ».

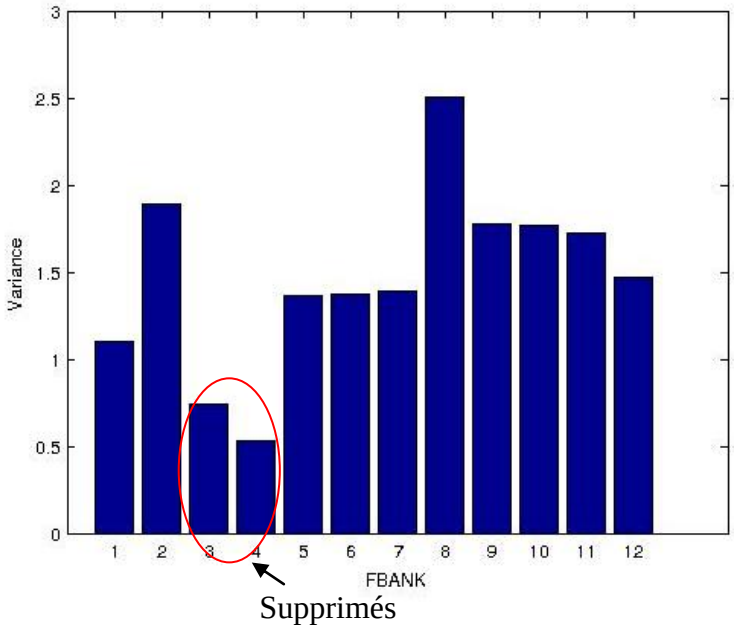

Figure 4. 9 – Variance des coefficients FBANK pour un exemple de 30 s de l'ensemble DEV du corpus « studio ».

La méthode consiste à examiner la variance du premier coefficient.

- Si elle est la plus faible, tous les coefficients à partir du second sont conservés. C'est le cas de l'exemple présenté dans la Figure 4. 8 (seul le premier coefficient est supprimé).

- Dans le cas contraire, les coefficients qui possèdent une variance inférieure à celle du premier coefficient sont supprimés. Ceci est illustré dans la Figure 4. 9 (les troisième et quatrième coefficients sont supprimés).

En utilisant cette méthode, la F-mesure est de 88,1% comme indiqué dans la dernière ligne du Tableau 4. 2. Ce qui correspond à des gains absolus de 9,4% et 0,6% par rapport aux résultats trouvés respectivement avec 24 FBANK et 12 FBANK. Le gain de la stratégie en termes de F-mesure n'est pas énorme par rapport au résultat trouvé avec 12 FBANK, mais nous avons privilégié cette méthode de paramétrisation car elle permet d'améliorer le rappel d'environ 10%. La précision a peu diminué mais elle reste satisfaisante et elle pourra augmenter lors de la phase de regroupement.

4.5.5. Influence du corpus de développement

Afin de tester l'influence du corpus de développement sur le choix du paramètre S_0 de la méthode DCAP et de consolider nos résultats plus profondément, nous avons distingué, dans nos enregistrements du corpus « studio », deux situations de tours de chant (cf. Figure 4. 10). La première correspond à l'introduction d'un silence entre deux chanteurs. La seconde contient une exacte alternance entre deux groupes de chanteurs. Un groupe de chanteurs peut être composé d'un seul chanteur (soliste) ou de plusieurs (chœurs). Plus de détails sur les enregistrements du corpus « studio » sont présentés dans le Chapitre 3.

Figure 4. 10 – Exemples de tours de chant rencontrés.

Il est clair que la transition chant-chant est la transition recherchée, l'autre pouvant être plus facilement trouvée par d'autres méthodes (comme une détection de silence ou une détection de chant…). Pour cette raison, nous avons divisé notre corpus en deux sous-corpus : le premier correspond uniquement aux situations de transition chant-chant (CC) et le deuxième contient les situations de transition chant-silence (CS). Chaque sous-corpus est découpé en un ensemble de développement (DEV) et un ensemble d'évaluation (EVAL). Nous nous sommes intéressés à l'étude du sous-corpus chant-chant et nous avons cherché à déterminer le paramètre S_0 sur ce sous-corpus (DEV-CC).

Le Tableau 4. 3 présente les résultats de la méthode DCAP avec la sélection de FBANK sur les deux sous-corpus. La meilleure performance en termes de F-mesure sur le DEV-CC est de 70,8% (valeur obtenue pour une valeur de $S_0 = 71$). A noter que cette valeur de S_0 est la

même que celle qui correspond à l'ensemble du DEV (cf. section 4.6). Par conséquent, nous pouvons déduire que même si le corpus de développement contient des situations (chant-silence) autres que celles correspondantes à des transitions chant-chant, la valeur du paramètre S_0 de la méthode DCAP ne varie pas et son ajustement ne dépend que des situations recherchées (chant-chant).

Tableau 4. 3 – Performances du système de segmentation avec la méthode DCAP sur les deux sous-corpus du corpus « studio ».

Méthode DCAP	**CC**			**CS**		
	Précision	**Rappel**	**F-mesure**	**Précision**	**Rappel**	**F-mesure**
DEV ($S_0 = 71$)	77	65,5	70,8	61,8	96,5	75,4
EVAL ($S_0 = 71$)	80	71,3	75,4	65	88,1	74,8

Les performances obtenues sur les ensembles d'évaluation du sous-corpus chant-chant (EVAL-CC) et du sous-corpus chant-silence (EVAL-CS) sont de 75,4% et 74,8% en termes de F-mesure. Les performances obtenues sont très similaires au DEV. Ces résultats montrent également que le système permet de bien segmenter les deux cas d'alternance.

4.6. Résultats globaux

Nous exposons dans cette section les résultats de notre système complet de segmentation en tours de chant sur le corpus « studio » avec les 12 MFCC et les FBANK sélectionnés.

Le Tableau 4. 4 présente les résultats de notre système de segmentation complet avec la méthode DCAP sur DEV et EVAL. Nous déterminons le paramètre S_0 de la méthode DCAP sur le DEV, qui a été aussi utilisé pour sélectionner les paramètres acoustiques. Dans cette méthode de segmentation, nous utilisons des matrices de covariance diagonales pour le modèle probabiliste du BIC. Le seuil S_0compris entre 1 et 161 est respectivement égal à 14, 68 et 71 avec les 12 MFCC, les 24 FBANK et les FBANK sélectionnés. Les résultats sont donnés avec une tolérance de 0,5 s.

Nous observons que la performance obtenue avec la sélection des coefficients FBANK est toujours meilleure que celle obtenue avec les 12 MFCC. Le gain absolu en termes de F-mesure est d'environ18,3% sur le DEV et de 30,6% sur l'EVAL (cf. Tableau 4. 4). Cela confirme les résultats obtenus avec le système « *oracle* » sur la partie DEV. De plus, les résultats trouvés avec la sélection des bandes restent meilleurs que ceux obtenus avec toutes les bandes (24 FBANK). Une amélioration de performance de 3,3% et 6,7% est réalisée respectivement sur le DEV et l'EVAL en utilisant les FBANK sélectionnés au lieu des 24 FBANK.

La grande différence de performance entre les MFCC et les FBANK, que ce soit avec toutes les bandes ou seulement les sélectionnées, se matérialise au niveau des fausses alarmes : le système en produit 1,5 et 2,2 fois plus avec les MFCC qu'avec les FBANK sélectionnés respectivement sur le DEV et l'EVAL. En analysant plus finement les résultats de segmentation, nous avons remarqué que les MFCC ont tendance à sur-segmenter en détectant les longues notes tenues plutôt que les tours de chant. Cela était aussi le cas lorsque nous utilisions les chromas.

Tableau 4. 4 – Résultats du système de segmentation avec DCAP sur le DEV et l'EVAL du corpus « studio ».

Paramètres	**DEV**			**EVAL**		
	Précision	**Rappel**	**F-mesure**	**Précision**	**Rappel**	**F-mesure**
12 MFCC	44	81,2	57	32,2	75,8	45,2
24 FBANK	70	74,1	72	67,3	71	69,1
FBANK sélectionnés	67,1	85,6	75,3	71,7	80,5	75,8

Par conséquent, les FBANK semblent être le type de paramètres acoustiques le plus adapté à notre contexte du chant. En effet, ils génèrent moins de fausses alarmes et permettent de détecter la plupart des frontières recherchées correspondantes aux points de changement de chanteurs : un rappel de 80,5% sur l'EVAL est obtenu.

Nous avons remarqué que les situations de transition soliste-chœur sont plus facile à détecter que les situations de transition soliste-soliste et chœur-chœur. Cela peut s'expliquer par le fait qu'en passant d'un soliste à un chœur et vice versa, le nombre de sources augmente et donc la détection devient plus facile.

4.7. Conclusion

Nous avons présenté un système de segmentation en tours de chant réalisé par analogie avec les systèmes de segmentation en tours de parole. Ce système se fonde sur le BIC, couramment utilisé pour la segmentation en tours de parole. Il est utilisé comme critère de décision de détection de frontières des segments de chant.

Nos contributions se situent dans trois domaines. Au niveau du paramètre de la taille de la fenêtre d'analyse du critère BIC, nous avons implémenté un algorithme de segmentation qui utilise une taille dynamique, inspiré de travaux réalisés pour la parole, et nous avons adaptés ces hyper-paramètres au contexte du chant.

Au niveau du second paramètre du BIC, le facteur de pénalité λ, nous avons proposé la méthode DCAP qui permet d'éviter le choix *a priori* de ce paramètre et d'améliorer la performance de notre système de segmentation.

Au niveau traitement du signal, nous avons proposé une nouvelle méthode de paramétrisation qui semble pertinente pour sélectionner les bandes fréquentielles les plus informatives pour notre tâche. Des comparaisons avec plusieurs paramètres acoustiques nous ont permis de valider le choix de la méthode de paramétrisation proposée.

L'application de notre système de segmentation sur les enregistrements de chant du corpus « studio » donne de bons résultats : une précision de 71,7%, un rappel de 80,5% et un score global de 75,8% en termes de F-mesure sur le corpus d'évaluation (EVAL). Ces résultats sont similaires aux résultats obtenus lorsque nous isolons les situations de transition chant-chant (CC) ou chant-silence (CS). A ce propos, une étape de prétraitement pourrait être ajoutée pour localiser les régions contenant du chant afin de limiter le traitement par la suite sur les situations de transition chant-chant, surtout lorsqu'il s'agit d'une grande quantité de données à traiter. Un autre prétraitement pourrait être la détection des zones des sons polyphoniques et monophoniques. Cela pourrait être effectué en tant que première étape de segmentation en soliste/chœur et serait considérée comme une initialisation pour le processus de segmentation en tours de chant.

En conclusion, les résultats de segmentation en tours de chant obtenus sur le corpus « studio », qui est de petite taille, sont néanmoins satisfaisants et encourageants.

Afin d'annoter les segments chantés par un même groupe de chanteur sous une même étiquette, une étape de regroupement doit être ajoutée pour obtenir un système complet de segmentation et regroupement en chanteurs (nom donné par analogie au système de segmentation et regroupement en locuteurs en parole). Le regroupement fera l'objet du chapitre suivant.

Chapitre 5

Regroupement en chanteurs

Sommaire

5.1. Introduction

Nous complétons l'étape de segmentation en tours de chant par une étape de regroupement en chanteurs qui fait l'objet de ce chapitre. Par analogie au regroupement en locuteurs qui permet d'identifier tous les segments prononcés par un même locuteur avec la même étiquette afin de répondre à la question « **qui parle et quand ?** », le regroupement en chanteurs vise à annoter tous les instants d'un enregistrement audio de chant avec des identifiants de classe pour répondre à la question « **qui chante et quand ?** ».

Plusieurs approches et critères existent en regroupement en locuteurs. Nous en avons étudié certains dans la partie 2.4.3. Parmi eux figure le critère BIC qui est très utilisé pour la segmentation en tours de parole et que nous avons adapté pour la segmentation en tours de chant. Nous étudions ce critère afin de réaliser notre système de regroupement en chanteurs.

Un système de base a été créé en utilisant l'outil de regroupement par BIC du LIUM. Nous avons commencé par appliquer ce système sur une segmentation parfaite afin d'évaluer le potentiel du critère BIC dans un contexte de chant. Ensuite, nous avons adapté ce système pour améliorer la performance du regroupement en chanteurs : nous avons ainsi proposé un module de fusion à la sortie du système de base. D'autres étapes ont été ajoutées après le module de fusion afin de peaufiner le regroupement en chanteurs.

L'intérêt des systèmes de regroupement que nous présentons réside dans le fait que nous n'utilisons pas un ajustement *a priori* du facteur de pénalité λ. En effet, la méthode de fusion fait suite à plusieurs regroupements par BIC.

Ce chapitre est divisé en quatre parties. Nous présentons, dans une première section, les approches de référence de regroupement que nous adaptons pour regrouper en chanteurs ainsi que les outils. Une deuxième section est consacrée à nos contributions et à la description des systèmes de regroupement en chanteurs proposés. Une étude de regroupement par BIC appliqué sur une segmentation parfaite fait l'objet de la troisième section. Les résultats finaux de nos deux systèmes de regroupement proposés sont présentés et interprétés dans la quatrième section.

5.2. Approches de référence de regroupement

Nous essayons d'effectuer le regroupement en chanteurs en empruntant des algorithmes de regroupement en locuteurs. La recherche en regroupement en locuteurs a commencé depuis environ 25 ans avec par exemple (Gish, et al., 1991) qui visait à rendre le regroupement robuste par rapport aux changements des conditions environnementales des locuteurs. Un autre problème apparaît lors de la modélisation des données quand les alternances entre locuteurs sont trop rapides. Le fait qu'il y ait eu beaucoup de travaux effectués sur le regroupement en parole, dont l'objectif et les facteurs de difficultés rencontrés se retrouveront pour le chant, nous a encouragé à procéder de la même manière pour regrouper en chanteurs.

Dans les approches les plus connues pour le regroupement en locuteurs, nous trouvons comme critère privilégié, le BIC. La méthode et le processus du déroulement du regroupement avec ce critère sont détaillés dans la partie 2.4.3.1 : il s'agit d'une approche de Classification Ascendante Hiérarchique (CAH). Elle prend souvent comme initialisation n classes, où n est le nombre de segments trouvés pendant la phase de segmentation. A chaque itération, les deux classes les plus proches au sens du critère ΔBIC sont fusionnées. L'utilisation du ΔBIC pour le regroupement ne dépend plus de la taille de la fenêtre comme pour la segmentation car en regroupement la taille de la fenêtre correspond à la taille des segments qui sont introduits à l'entrée de ce processus, mais elle dépend encore de la valeur du facteur de pénalité λ. Généralement pour le regroupement en locuteurs, la valeur de ce paramètre est déterminée sur un ensemble de développement.

La plupart des systèmes de segmentation et regroupement en locuteurs les plus performants utilisent une première passe principale de regroupement par BIC et la complète par la suite par d'autres passes supplémentaires en utilisant d'autres critères tels que le CE, le CLR et l'ILP/i-vectors. Nous avons détaillé ces critères dans la partie 2.4.3. Deux exemples de système « complet » de l'IRIT et du LIUM de segmentation et regroupement en locuteurs sont aussi décrits dans la partie 2.4.4. Profitant de la disponibilité du système de segmentation et regroupement en locuteurs de LIUM[6], nous utilisons les outils de regroupement de la bibliothèque *LIUM_SpkDiarization*. Cette bibliothèque a été développée dans le cadre de la campagne d'évaluation ESTER2 en 2008 (Galliano, et al., 2009).

5.3. Contributions : présentation de nos systèmes de regroupement en chanteurs

Nous présentons dans cette section nos différentes contributions pour le regroupement en chanteurs. Celles-ci sont divisées en trois parties :

- la première contribution consiste à réaliser un système de regroupement de base obtenu en utilisant notre paramétrisation au sein du système de référence de regroupement par BIC,
- la deuxième contribution est la proposition d'un système de regroupement que nous appelons RDCAP obtenu en ajoutant une étape de fusion à la sortie du système de base,
- et la troisième contribution consiste à ajouter des passes supplémentaires de regroupement au système RDCAP, ce qui conduit à avoir un autre système, nommé RDCAP+VCE.

Nous tenons à préciser une nouvelle fois que le système de référence de regroupement par le critère BIC utilisé dans cette thèse est celui développé par le LIUM.

[6] http://www-lium.univ-lemans.fr/diarization/doku.php/welcome

5.3.1. Système de regroupement en chanteurs de base

Le système de regroupement en chanteurs de base est constitué de deux étapes. La première est la paramétrisation à l'aide des mêmes coefficients FBANK que ceux déjà utilisés lors de la phase de segmentation en tours de chant. Nous rappelons que cette méthode de paramétrisation repose sur une stratégie de sélection des bandes fréquentielles pertinentes en choisissant les coefficients FBANK qui possèdent une variance importante et en s'arrêtant à la douzième bande. La seconde étape est l'application du regroupement par BIC. Comme dit dans le paragraphe 5.2, pour ce système le regroupement par BIC dépend encore du paramètre de pénalité λ qui est fixé *a priori* sur un ensemble de développement : DEV du corpus « studio ».

L'ajustement *a priori* de la valeur de λ pour l'étape de regroupement pose le même problème que celui rencontré en segmentation en tours de chant : la valeur optimale de λ varie grandement d'un enregistrement à un autre. En effet, le type de chant et les notes (courtes ou tenues) peuvent être très variables d'un enregistrement à un autre (même au sein du même corpus) et d'un groupe de chanteurs à un autre dans un même enregistrement. Ce problème ne se pose pas en parole car le débit syllabique est assez constant, de l'ordre de 4 Hertz. Par contre le débit en chant peut varier d'une phrase chantée à une autre (changement de notes) et d'un chanteur à un autre. Cela implique, que le choix d'une unique valeur dite « optimale » pour λ n'est pas raisonnable dans un contexte de chant.

5.3.2. Système de regroupement en chanteurs RDCAP

Afin de pallier au problème de variabilité du facteur de pénalité λ et d'éviter le choix *a priori* de sa valeur, nous avons procédé de la même manière que pour la segmentation en tours de chant et nous avons décidé d'utiliser une méthode de fusion pour le regroupement, que nous appelons Regroupement avec Décision par Consolidation *A Posteriori* (RDCAP). La méthode RDCAP que nous avons proposée consiste à concaténer les sorties de plusieurs systèmes de regroupement par BIC, obtenues en faisant varier la valeur du coefficient de pénalité λ sur un intervalle de valeurs, ajusté sur un ensemble de développement. Pour décider quelles sont les étiquettes de classes à considérer, un vote majoritaire est effectué à partir de la concaténation de toutes les sorties des systèmes : chaque segment du chant est identifié par l'étiquette de classe choisie par la majorité des systèmes de regroupement obtenus. La proposition de cette méthode implique la réalisation d'un nouveau système de regroupement en chanteurs que nous appelons RDCAP. Son architecture est illustrée dans la Figure 5. 1.

Ce système est composé de trois étapes :

1. La première permet d'extraire les paramètres acoustiques à utiliser qui sont les FBANK sélectionnés complétés par l'énergie.
2. La deuxième réalise plusieurs regroupements par le critère BIC en faisant varier à chaque fois la valeur du coefficient de pénalité λ.
3. La troisième consiste à appliquer la méthode de fusion RDCAP en concaténant, tout d'abord, toutes les sorties des systèmes de regroupement obtenus pendant l'étape précédente et en effectuant, par la suite, le vote majoritaire.

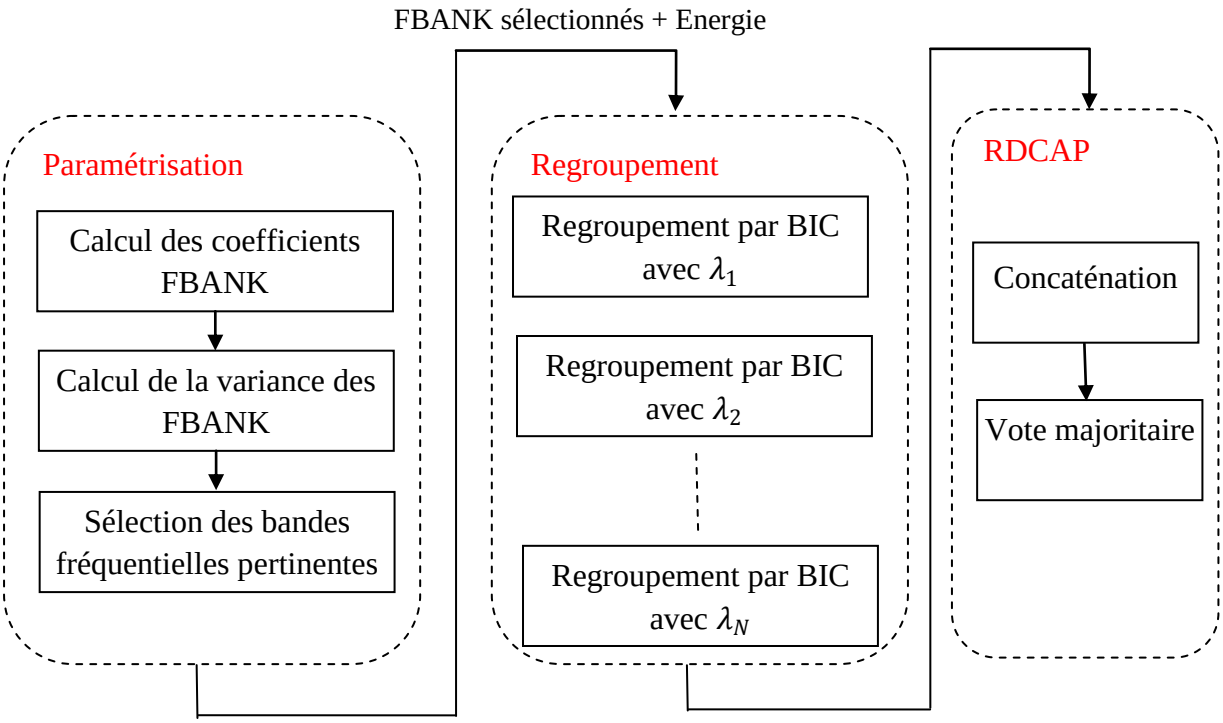

Figure 5. 1 – Architecture du système de regroupement en chanteurs RDCAP.

5.3.3. Système de regroupement en chanteurs RDCAP+VCE

Nous proposons un autre système de regroupement en chanteurs qui complète le système RDCAP par d'autres modules supplémentaires qui sont le Viterbi et l'Entropie Croisée (Cross Entropy – CE). Nous appelons ce nouveau système RDCAP+VCE et nous illustrons son architecture dans la Figure 5. 2.

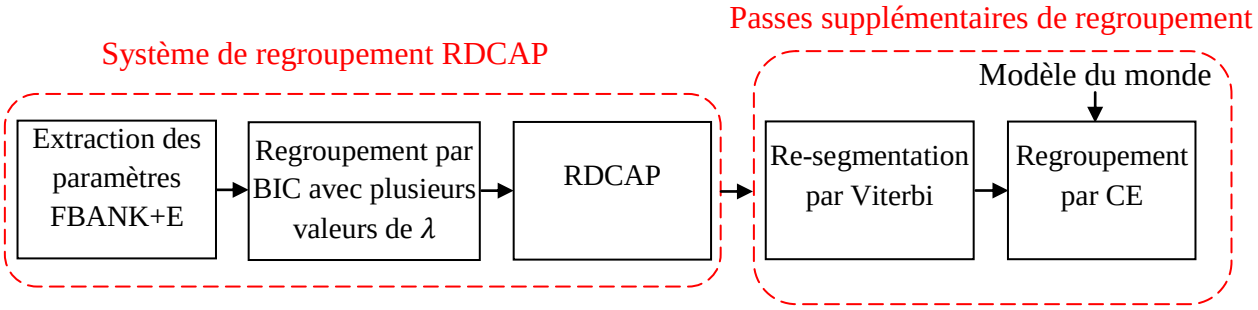

Figure 5. 2 – Architecture du système de regroupement en chanteurs RDCAP+VCE.

Ce système est composé de cinq étapes :

1. La première permet d'extraire les paramètres acoustiques FBANK sélectionnés et l'énergie,
2. La deuxième consiste à réaliser plusieurs regroupements BIC avec différentes valeurs de λ,
3. La troisième applique la méthode de fusion RDCAP sur les sorties de l'étape précédente,

4. La quatrième consiste à réaliser un ajustement des frontières des segments par une re-segmentation. Pour cela, un GMM de 8 gaussiennes est appris pour chaque classe en utilisant tous les segments de la classe. Les modèles sont obtenus en utilisant l'algorithme Expectation Maximisation (EM) et une nouvelle segmentation est obtenue après application de l'algorithme de Viterbi. Un rapport de vraisemblance est calculé entre les classes initiales (résultantes du regroupement avec BIC et RDCAP) et les classes résultant du raffinement des frontières par Viterbi. Plusieurs itérations sont effectuées : à chaque itération, les frontières des segments de chaque classe sont affinées jusqu'à ce que le rapport de vraisemblance soit inférieur à un certain seuil,
5. La cinquième étape effectue un regroupement par le CE. Cette étape finale de regroupement nécessite la construction d'un modèle du monde (Universal Background Model – UBM) qui est appris à partir d'un ensemble d'apprentissage composé de 8 gaussiennes avec l'algorithme EM. Ce modèle est adapté, par la suite, à chaque classe pour obtenir le modèle de groupe de chanteur relatif à cette classe. A chaque itération, les deux classes qui maximisent la mesure CE sont fusionnées jusqu'à ce que cette mesure dépasse un certain seuil.

Nous utilisons là aussi les modules de la boîte à outils *LIUM_SpkDiarization* pour réaliser les étapes de re-segmentation par Viterbi et de regroupement par CE.

5.4. Etude du regroupement par BIC avec une segmentation parfaite

Une première expérimentation de regroupement en chanteurs est faite à partir d'une segmentation parfaite ; elle servira de référence afin de s'assurer des résultats des outils appliqués sur le chant, c'est-à-dire des erreurs liées au regroupement indépendamment des erreurs engendrées par la segmentation. Tous les résultats présentés dans cette section sont obtenus en utilisant le corpus « studio » et les segmentations de référence des enregistrements de ce corpus.

Dans cette section, nous présentons, dans une première partie, les expériences initiales que nous avons réalisées ainsi que le système « *oracle* » du regroupement. La deuxième partie est consacrée à l'évaluation du système de regroupement de base proposé et une étude par fichier est illustrée. Dans la troisième partie, nous parlons de l'application du système RDCAP ainsi que de la possibilité de proposer une stratégie pour déterminer automatiquement l'intervalle optimal de la valeur de λ pour chaque enregistrement.

5.4.1. Premières expériences et système « *oracle* »

Nous avons débuté nos expériences en appliquant un regroupement par BIC sur nos enregistrements de chant du corpus « studio ». Nous avons utilisé les paramètres FBANK et l'énergie. Nous avons essayé plusieurs valeurs du facteur de pénalité du BIC, en le faisant varier sur l'intervalle $[0{,}5 \quad 12{,}0]$ avec un pas de 0,5. Chaque classe est modélisée par une gaussienne avec une matrice de covariance pleine. Nous avons remarqué que les performances étaient stables sur plusieurs sous-intervalles de λ, ce qui nous a conduit à ne pas

utiliser des valeurs de pas inférieures à 0,5. Les performances du regroupement sont évaluées avec le *Diarization Error Rate* (DER), qui est utilisé pour l'évaluation des systèmes de regroupement en locuteurs, et dont la méthode de calcul est détaillée dans la partie 3.5.2.

Nous avons décidé de calculer la performance du système « *oracle* » de regroupement pour avoir une idée du taux d'erreur (DER) le plus faible que nous pouvons avoir. Pour cela, nous avons déterminé pour chaque enregistrement des ensembles DEV et EVAL la valeur optimale de λ qui permet d'avoir le meilleur DER (le plus petit). Les résultats du système « *oracle* » avec la segmentation parfaite (manuelle) sont illustrés dans le Tableau 5. 1. Nous illustrons le DER pour le DEV et l'EVAL avec trois tolérances différentes sur les frontières : 0 s, 0,25 s et 0,5 s. En parole, une tolérance de 0,25 s est souvent utilisée dans le cas des enregistrements contenant de la parole préparée. Lorsqu'il s'agit de parole spontanée, une tolérance plus faible est utilisée (Meignier, 2015). Dans notre contexte de musique, nous considérons une tolérance de 0,5 s, la même que nous avions utilisée lors de l'évaluation de l'étape de segmentation en tours de chant.

Nous obtenons un taux d'erreur (DER) de 10,38% sur le DEV et 12,68% sur l'EVAL pour une tolérance de 0,5 s. Nous trouvons un DER de 14,11% et 15,73% avec une tolérance plus faible (0,25 s) et nulle sur l'EVAL. Ces taux d'erreurs ne sont pas élevés en comparaison de ceux obtenus en parole : ceci nous rassure et nous encourage à continuer à utiliser le critère BIC pour le regroupement dans un contexte de chant.

Tableau 5. 1 – Résultats du système « oracle » de regroupement, suite à une segmentation manuelle des ensembles DEV et EVAL du corpus « studio ».

Tolérance	DEV DER	EVAL DER
0,50 s	10,38%	12,68%
0,25 s	12,57%	14,11%
0,00 s	14,67%	15,73%

5.4.2. Application du système de base

Après avoir obtenu des résultats encourageants pour le système « *oracle* », nous avons testé notre système de regroupement de base et ajusté la valeur de λ sur le DEV. Nous avons fait varier la valeur du facteur de pénalité sur le même intervalle que celui utilisé pour l'expérience précédente du système « *oracle* ». Les résultats de ce système avec une segmentation manuelle sont illustrés dans le Tableau 5. 2. La valeur optimale de λ que nous avons déterminée sur le DEV est égale à 10,5. Pour cette valeur, nous obtenons 12,90% d'erreurs sur le DEV et 18,47% d'erreurs sur l'EVAL avec une tolérance de 0,5 s. Si nous considérons une tolérance plus faible (0,25 s) et une tolérance nulle, le DER augmente de 3,94% et 7,70% respectivement. Naturellement, le taux d'erreurs reste plus élevé que celui du système « *oracle* » : il existe encore une marge de réduction d'erreurs possible d'environ 6%

sur l'EVAL, si nous considérons le DER obtenu avec le système « *oracle* » comme la limite inférieure à atteindre.

Tableau 5. 2 – Résultats du système de regroupement de base, suite à une segmentation manuelle des ensembles DEV et EVAL du corpus « studio ».

Tolérance	**DEV DER**	**EVAL DER**
0,50 s	$\lambda = 10{,}5$; 12,90%	18,47%
0,25 s	$\lambda = 10{,}5$; 16,27%	22,41%
0,00 s	$\lambda = 10{,}5$; 19,43%	26,17%

Nous avons analysé les scores par fichier pour repérer la raison de l'augmentation de l'erreur avec le système de base par rapport au système « *oracle* ». Nous présentons dans le Tableau 5. 3 la durée de chaque fichier de l'ensemble de développement (DEV) et de l'ensemble d'évaluation (EVAL) ainsi que les meilleures valeurs de λ associées (les valeurs de λ du système « *oracle* »). Dans ce tableau se trouvent également les DER du système « *oracle* » et du système de base (avec une valeur fixe de $\lambda = 10{,}5$) avec une tolérance de 0,5 s obtenus avec la segmentation manuelle. Les fichiers du DEV portent le suffixe « _dev » et ceux de l'EVAL portent le suffixe « _eval ».

Tableau 5. 3 – Valeurs de λ ainsi que le DER par fichier du système « oracle » et le DER par fichier avec le système de regroupement de base sur les ensembles DEV et EVAL du corpus « studio », suite à une segmentation manuelle.

Nom des fichiers	**Durée(s)**	**λ « *Oracle* »**	**DER «*Oracle*»**	**DER ($\lambda = 10{,}5$)**
03-Mayingo_dev	31	[3,0 5,5]	2,05	7,48
Arranoak_Bortietan_dev	58	[10,5 12,0]	26,54	26,54
Malicorne_Marion_Les_Roses_dev	69	[8,5 12,0]	0,0	0,0
sloopJohnB_dev	38	7,0	13,06	24,93
03-Mayingo_eval	45	[1,5 5,0]	0,0	8,51
Arranoak_Bortietan_eval	97	9,0	46,35	46,90
Malicorne_Marion_Les_Roses_eval	141	[8,0 12,0]	0,87	0,87
sloopJohnB_1_eval	57	[6,0 6,5]	7,94	56,08
sloopJohnB_2_eval	95	[5,5 8,5]	2,87	4,00

Remarques : les fichiers « 03-Mayingo » et « sloopJohnB » contiennent des alternances entre différents groupes de chanteurs sans alternance avec du silence. Les fichiers « Malicorne_Marion_Les_Roses » et « Arranoak_Bortietan » contiennent des zones de

silence. Les fichiers de « Malicorne_Marion_Les_Roses » contiennent deux groupes de chanteurs (deux chœurs) alternés avec du silence. Les fichiers « Arranoak_Bortietan » contiennent un seul chanteur (soliste) qui fait des pauses en chantant.

A partir du Tableau 5. 3, nous remarquons que la meilleure valeur de λ varie selon les fichiers et que la valeur de 10,5 du système de base n'est pas la bonne pour tous les fichiers du DEV. En effet, en utilisant une valeur fixe du facteur de pénalité, le DER augmente d'environ 5,5% pour le fichier « 03-Mayingo_dev » et presque de 12% pour le fichier « sloopJohnB_dev ». Pour l'EVAL l'écart est encore plus important : pour le fichier « sloopJohnB_1_eval », le DER passe de 7,94% avec le système « *oracle* » à 56,08% en utilisant la valeur fixe de 10,5 pour λ. Pour ce dernier, toutes les classes ont été regroupées dans une seule alors qu'il en existe trois en vérité : une classe pour un chœur composé de deux chanteurs, une classe de non-chant qui contient une faible réverbération, et une autre classe de chœur composé de trois chanteurs dont les deux chanteurs du premier chœur sont présents. La présence des chanteurs en commun dans les deux chœurs semble expliquer ce phénomène de sur-regroupement. Pour ne pas avoir ce phénomène, il faut prendre une valeur de λ plus faible. En effet, il s'agit d'un comportement naturel du BIC en fonction de la valeur du coefficient de pénalité : en augmentant sa valeur, nous obtenons un sur-regroupement et en diminuant sa valeur nous obtenons un sous-regroupement.

5.4.3. Application du système RDCAP

Afin de réduire le problème de variabilité du facteur de pénalité λ, nous utilisons la méthode de fusion RDCAP (cf. partie 5.3). Comme nous pouvons le remarquer dans le Tableau 5. 3, les meilleures valeurs de λ pour chaque fichier ainsi que la valeur fixée sur la totalité du DEV sont incluses dans l'intervalle $[5{,}0\ 12{,}0]$. Pour cela, nous avons décidé d'utiliser cet intervalle dans lequel nous faisons varier la valeur de λ avec un pas de 0,5. Ainsi, 15 systèmes de regroupement sont obtenus et concaténés. Les résultats de l'application du système RDCAP avec la segmentation manuelle sur les ensembles DEV et EVAL sont présentés dans le Tableau 5. 4.

Tableau 5. 4 – DER du système de regroupement RDCAP, suite à une segmentation manuelle sur les ensembles DEV et EVAL du corpus « studio ».

Tolérance	DEV DER	EVAL DER
0,50 s	19,68%	16,66%
0,25 s	21,80%	20,49%
0,00 s	23,77%	24,15%

En appliquant le système RDCAP, nous trouvons un DER de 16,66% sur l'EVAL, soit un gain d'environ 2% par rapport au système de base qui utilise une valeur fixe de λ (cf. Tableau 5. 2). Ce résultat est encourageant mais il reste encore une marge de 4% pour atteindre la borne inférieure du taux d'erreurs qui est de 12,68% sur l'EVAL.

En regardant les scores par fichier sur l'EVAL, nous avons trouvé que le DER pour le fichier « Arranoak_Bortietan_eval » reste toujours élevé : environ 48%. Toutes les zones de silence sont regroupées dans une seule classe, par contre, les zones chantées par le même soliste sont classées en plusieurs classes. Ce problème de sous-regroupement est lié à la difficulté de ce fichier qui contient plusieurs notes tenues qui faussent le processus de regroupement en n'arrivant pas à identifier les classes du même chanteur par la même étiquette. Ce phénomène nous a conduit à étudier le nombre de classes trouvées en fonction de la valeur de λ. Nous avons tracé la courbe du nombre de classes en fonction du facteur de pénalité λ pour chaque fichier en faisant varier sa valeur dans l'intervalle initial [0,5 12,0] avec un pas de 0,5 (cf. Figure 5. 3 pour l'ensemble DEV et Figure 5. 4 pour l'ensemble EVAL).

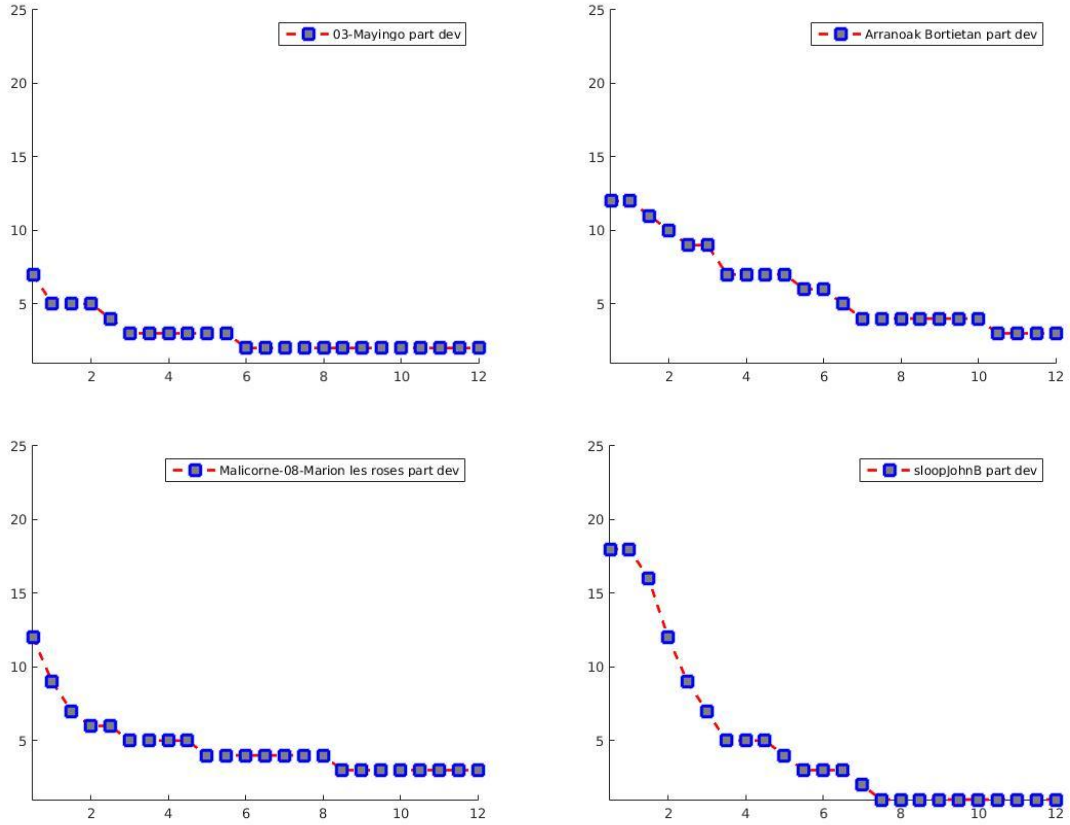

Figure 5. 3 – Courbes du nombre de classes de groupes de chanteur(s) en fonction de la valeur de λ pour chaque fichier de l'ensemble DEV.

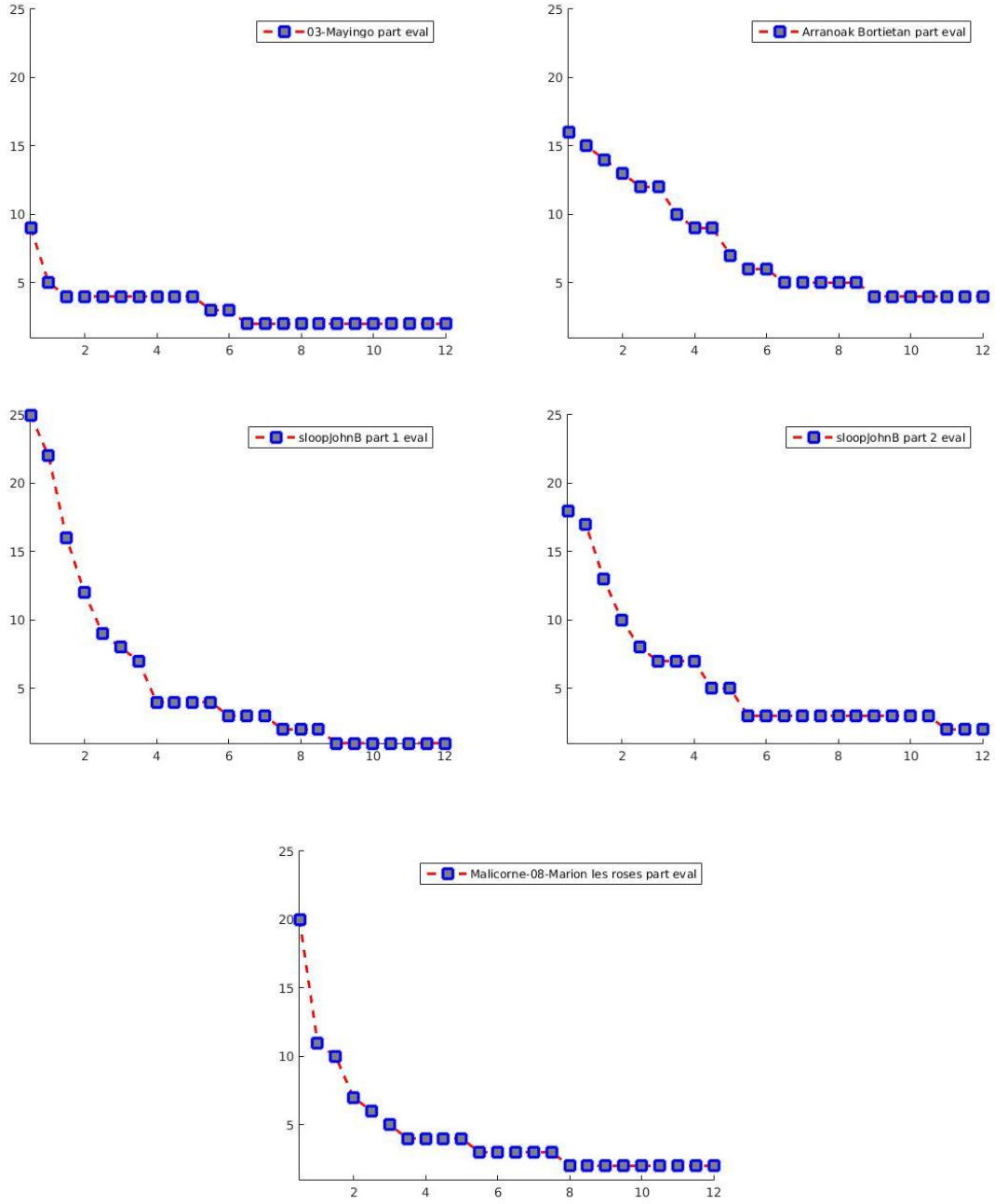

Figure 5. 4 – Courbes du nombre de classes de groupes de chanteur(s) en fonction de la valeur de λ pour chaque fichier de l'ensemble EVAL.

Nous remarquons sur ces courbes la présence de « plateaux » pour lesquels le nombre de classes ne change pas pour un ensemble de valeurs de λ successives. A partir de cette visualisation, nous avons mis en place une stratégie pour choisir automatiquement l'intervalle de λ sur lequel nous appliquons notre système RDCAP. Cette stratégie consiste tout d'abord à choisir le plus petit des deux derniers plateaux. Puis, nous considérons les valeurs de λ correspondant au plateau choisi et nous les complétons avec des valeurs des deux cotés du plateau jusqu'à arriver à 9 valeurs. Nous illustrons cette stratégie dans la Figure 5. 5 sur deux exemples du DEV en encadrant en rouge l'intervalle des valeurs de λ choisi automatiquement.

Les intervalles automatiques, les résultats obtenus avec cette stratégie ainsi que les résultats trouvés avec un intervalle fixe ([5,0 12,0]) pour chaque fichier avec le système RDCAP, suite à une segmentation manuelle, sont présentés dans le Tableau 5. 5.

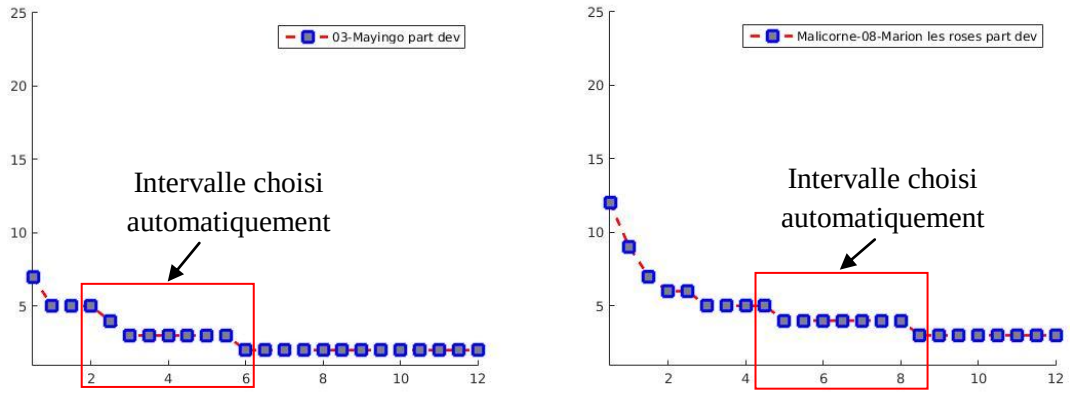

Figure 5. 5 – Illustration de la stratégie de choix automatique d'intervalles de variation de λ sur deux exemples de l'ensemble DEV. La partie cadrée en rouge sur chaque exemple correspond à l'intervalle de valeurs de λ choisi automatiquement.

Tableau 5. 5 – Intervalles automatiques des valeurs de λ et les résultats obtenus avec la stratégie ainsi que les résultats trouvés avec un intervalle fixe du système RDCAP pour chaque fichier des ensembles DEV et EVAL du corpus « studio », suite à une segmentation manuelle.

Nom des fichiers	**λ – Intervalle automatique**	**DER – RDCAP Intervalle automatique**	**DER – RDCAP Intervalle fixe [5,0 12,0]**
03-Mayingo_dev	[2,0 6,0]	2,05	7,48
Arranoak_Bortietan_dev	[9,0 9,5]	26,54	48,39
Malicorne_Marion_Les_Roses_dev	[4,5 8,5]	34,06	0,0
sloopJohnB_dev	[4,0 8,0]	27,62	24,93
03-Mayingo_eval	[3,5 7,5]	3,53	8,51
Arranoak_Bortietan_eval	[5,5 9,5]	48,43	48,43
Malicorne_Marion_Les_Roses_eval	[4,5 8,5]	35,02	0,87
sloopJohnB_1_eval	[6,0 10,0]	33,16	33,16
sloopJohnB_2_eval	[9,5 13,5]	4,00	4,00

La stratégie a bien marché sur 6 fichiers et elle a échoué pour 3 qui sont : « sloopJohnB_dev » pour lequel une augmentation en DER de 2,69% est obtenue, « Malicorne_Marion_Les_Roses_dev » et « Malicorne_Marion_Les_Roses_eval » pour les quels les taux d'erreurs ont augmenté d'environ 34% en comparant les résultats obtenus avec cette stratégie par rapport à ceux trouvés avec un intervalle fixe. L'importante augmentation

trouvée avec les deux fichiers de « Malicorne_Marion_Les_Roses » est dû au fait que les intervalles choisis automatiquement ne contiennent pas assez de valeurs de λ qui donnent la bonne performance. Ce qui fait que l'étape de vote majoritaire est faussée puisque la plupart des sorties des systèmes considérés sont obtenus avec des valeurs non optimales de λ. Par conséquent, nous n'avons pas retenu cette stratégie pour la suite et nous sommes restés sur le choix de l'intervalle $[5{,}0 \;\; 12{,}0]$ qui nous permet de réduire le taux d'erreurs de 2% en absolu par rapport au système de base sur l'EVAL.

5.5. Evaluation des systèmes de regroupement RDCAP et RDCAP+VCE avec notre segmentation automatique

Après avoir validé la méthode de regroupement par BIC et RDCAP avec une segmentation parfaite (manuelle), nous l'avons appliquée sur notre segmentation automatique obtenue avec notre système de segmentation en tours de chant décrit dans le Chapitre 4.

Dans cette section, nous présentons, dans une première partie, l'évaluation de notre système de regroupement RDCAP, constitué des deux étapes BIC et RDCAP. La deuxième partie est consacrée à l'évaluation du système de regroupement RDCAP+VCE, composé du BIC, RDCAP, re-segmentation par Viterbi et d'une dernière étape de regroupement par CE.

5.5.1. Evaluation du système RDCAP

Afin d'évaluer notre système de regroupement RDCAP en conditions réelles (suite à une segmentation automatique), nous avons commencé par évaluer le système « *oracle* » en faisant varier le facteur de pénalité λ dans l'intervalle $[5{,}0 \;\; 12{,}0]$. Les résultats sont illustrés dans le Tableau 5. 6 avec trois tolérances. Nous remarquons que le taux d'erreur obtenu sur l'EVAL (13,60%) n'est pas très élevé par rapport à celui obtenu avec une segmentation manuelle (12,68%, cf. Tableau 5. 1). Cela semble prouver que notre segmentation automatique est performante : elle n'engendre que très peu d'erreurs.

Tableau 5. 6 – DER du système de regroupement « oracle », suite à la segmentation automatique sur les ensembles DEV et EVAL du corpus « studio ».

Tolérance	DEV DER	EVAL DER
0,50 s	20,37%	13,60%
0,25 s	24,16%	16,26%
0,00 s	28,42%	20,34%

En appliquant le système de regroupement RDCAP sur notre segmentation automatique, nous obtenons un DER de 19,7% (resp. 23,05% et 27,46%) sur l'EVAL avec une tolérance de 0,5 s (resp. 0,25 s et 0 s) (cf. Tableau 5. 7).

Tableau 5. 7 – DER du système RDCAP, suite à la segmentation automatique sur les ensembles DEV et EVAL.

Tolérance	DEV DER	EVAL DER
0,50 s	27,07%	**19,70%**
0,25 s	28,96%	23,05%
0,00 s	31,51%	27,46%

5.5.2. Evaluation du système RDCAP+VCE

Afin de réaliser le système RDCAP+VCE, nous avons ajouté des passes supplémentaires au système RDCAP. Nous avons appris un modèle du monde (UBM) sur les données de développement (DEV) pour la réalisation du regroupement par le critère CE.

L'application du système RDCAP+VCE, dont l'architecture est décrite dans la section 5.3.3, nécessite l'ajustement d'un seuil δ pour la dernière étape de regroupement, à partir duquel le processus de fusion avec la métrique CE s'arrête. Pour cela, nous utilisons également l'ensemble DEV. Nous faisons varier le seuil entre les valeurs 0,5 et 2,5 avec un pas de 0,5. La valeur de δ qui nous permet d'avoir le DER le moins élevé sur le DEV est égale à 1,0. Les scores trouvés avec cette valeur sont présentés dans le Tableau 5. 8.

Tableau 5. 8 – DER du système RDCAP+VCE, suite à la segmentation automatique sur les ensembles DEV et EVAL.

Tolérance	DEV DER	EVAL DER
0,50 s	16,21%	**10,91%**
0,25 s	19,70%	15,79%
0,00 s	23,77%	20,85%

Nous obtenons un DER de 10,91% (resp. 15,79% et 20,85%) pour une tolérance de 0,5 s (resp. 0,25 s et 0 s) sur l'EVAL avec le système RDCAP+VCE : nous avons réduit le taux d'erreurs d'environ 9% par rapport au système RDCAP. Ainsi, ce taux est désormais comparable à ceux obtenus par les meilleurs systèmes de regroupement en locuteurs.

Afin de confirmer la valeur ajoutée de notre méthode de fusion RDCAP, nous avons testé toutes les passes de regroupement du système RDCAP+VCE sans le module RDCAP, ce qui revient à appliquer le système de regroupement du LIUM, décrit dans la partie 2.4.4.2, avec notre paramétrisation FBANK à la place des MFCC. Nous avons ajusté la valeur de λ du BIC ainsi que le seuil δ de la CE en considérant le couple de valeurs (λ, δ) qui nous permet d'avoir le DER le moins élevé sur l'ensemble DEV, sachant que nous faisons varier λ et δ dans les mêmes intervalles que ceux utilisés précédemment pour le système RDCAP+VCE. Le meilleur score trouvé sur le DEV est obtenu avec une valeur de δ égale à 1,0 pour

n'importe quelle valeur de λ appartenant au sous-intervalle $[10{,}5 \;\; 11{,}5]$. Les résultats trouvés sont illustrés dans le Tableau 5. 9.

Tableau 5. 9 – DER du système RDCAP+VCE, suite à la segmentation automatique sur les ensembles DEV et EVAL.

Tolérance	**DEV DER**	**EVAL DER**
0,50 s	15,00%	**20,52%**
0,25 s	18,64%	24,59%
0,00 s	22,88%	28,90%

En utilisant le système de regroupement de LIUM, nous obtenons un DER de 20,52% (resp. 24,59% et 28,90%) avec une tolérance de 0,5 s (resp. 0,25 s et 0 s) sur l'EVAL, qui est presque 2 fois plus élevé que le taux d'erreurs obtenu avec notre système RDCAP+VCE. Cela semble prouver l'utilité de notre méthode RDCAP dans un contexte de chant.

Afin de pouvoir comparer clairement les résultats de regroupement obtenus suite à notre segmentation automatique, nous avons listé tous les résultats dans le Tableau 5. 10. Nous observons qu'avec un regroupement par BIC et notre méthode de fusion seulement, nous obtenons une performance meilleure que celle trouvée avec le système LIUM qui effectue des passes supplémentaires après le regroupement par BIC : un DER de 19,70% avec le système RDCAP contre un DER de 20,52% avec le système LIUM. En utilisant notre système RDCAP+VCE effectuant des passes supplémentaires après le BIC et notre méthode de fusion, nous améliorons encore le résultat de 9,61% et 8,79% par rapport aux systèmes LIUM et RDCAP.

Tableau 5. 10 – Résultats des différents systèmes de regroupement testés, suite à la segmentation automatique sur les ensembles DEV et EVAL.

Système	**DEV DER**	**EVAL DER**
RDCAP	27,07%	19,70%
LIUM	15,00%	20,52%
RDCAP+VCE	16,21%	10,91%

5.6. Conclusion

Afin de construire notre système de regroupements en chanteurs, nous nous sommes appuyés sur des approches de regroupement en locuteurs. Nous avons ainsi utilisé des outils de regroupement développés par le LIUM.

Nos contributions sont diverses. Au niveau de la paramétrisation, comme lors de la segmentation en chanteurs, nous avons utilisé notre méthode de sélection de bandes fréquentielles pertinentes (FBANK). Au niveau du regroupement, nous avons proposé la méthode RDCAP par analogie à la méthode DCAP que nous avions implémentée pour la segmentation en tours de chant. Cette méthode effectue, tout d'abord, une concaténation des sorties de plusieurs systèmes de regroupement par BIC, ensuite, elle réalise un vote majoritaire pour choisir, pour chaque segment, la classe à laquelle il va être attribué.

Afin de valider la robustesse de ce système et d'ajuster son paramètre (intervalle de variation de λ), nous l'avons testé, au départ, sur une segmentation manuelle. Sur celle-ci, nous avons obtenu un taux d'erreurs de 16.6% sur l'EVAL, soit 2% de mieux que le système de base de regroupement par BIC qui utilise une valeur fixe de λ déterminée sur le corpus de développement.

Ensuite, nous avons appliqué ce système sur une segmentation automatique, sortie de notre système de segmentation en tours de chant du chapitre précédent. Nous avons obtenu un DER d'environ 19%.

Puis, nous avons proposé de compléter le système RDCAP par d'autres passes de regroupement (Viterbi+CE), classiquement utilisées en regroupement en locuteurs. Cet ensemble du système RDCAP et de ces étapes de regroupement (BIC+RDCAP+Viterbi+CE) constitue un deuxième système de regroupement en chanteurs appelé RDCAP+VCE. L'application de ce système sur notre segmentation automatique nous donne un taux d'erreurs de 10,9%, qui est deux fois moins élevé que celui obtenu avec le système RDCAP ou avec le système de regroupement du LIUM.

Par conséquent, les deux systèmes proposés RDCAP et RDCAP+VCE ont prouvé leur intérêt sur une tâche de regroupement en chanteurs sur le corpus « studio ». Leurs robustesses par rapport aux conditions d'enregistrement seront traitées dans le chapitre suivant qui présente l'application du système complet de segmentation et regroupement en chanteurs sur des enregistrements ethnomusicologiques hétérogènes (corpus DIADEMS).

Chapitre 6

Application de notre système de segmentation et regroupement en chanteurs sur des enregistrements ethnomusicologiques du projet DIADEMS

Sommaire

6.1. Introduction

L'étude d'enregistrements ethnomusicologiques, dans le cadre du projet ANR CONTINT DIADEMS dont l'IRIT est le porteur, est à l'origine de notre étude de segmentation et regroupement en chanteurs. Ce projet est composé de huit partenaires : quatre partenaires en Sciences et Technologie de l'Information et de Communication (STIC) qui sont l'IRIT, le LaBRI, le LIMSI et le LAM, trois partenaires en ethnomusicologie qui sont le LESC, le CREM et la MNHN et un partenaire industriel, la société Parisson.

Les archives du Laboratoire d'Ethnologie et de Sociologie Comparative (LESC), le fonds du Centre de Recherche en Ethnomusicologie (CREM) et du Laboratoire d'Eco-anthropologie et Ethnobiologie du Muséum National d'Histoire Naturelle (MNHN) définissent une collection d'une grande importance historique et unique au monde, de plus de 4000 heures de musiques inédites. Ces enregistrements viennent de plusieurs pays de l'Afrique, Europe, Asie, Amériques, etc. L'enjeu du projet est de rendre ce patrimoine immatériel mondial accessible et exploitable scientifiquement grâce à des outils modernes et innovants dans le secteur de l'audiovisuel, plus particulièrement dans le traitement du son. Son objectif est d'étudier et implémenter diverses technologies du traitement automatique du son afin de faciliter l'accès à ces archives sonores. Ce qui revient à réaliser des outils de structuration et d'indexation d'un document sonore ou d'une collection de documents. Plus de détails sur le projet DIADEMS ainsi que les différentes tâches demandées sont disponibles sur le site Web[7] du projet, que nous avons développé pendant la phase de gestion du projet dont l'IRIT est responsable.

Dans ce contexte d'indexation de documents ethnomusicologiques sonores, le repérage des chanteurs et des chœurs est apparu comme essentiel et nous a amené à s'interroger sur la notion de « tours de chant ». Après avoir développé un système de segmentation et regroupement en chanteurs et l'avoir validé sur un corpus enregistré dans des conditions studio (corpus « studio »), nous étudions dans ce chapitre l'application de ce système sur le corpus appelé corpus DIADEMS.

Plusieurs tâches ont été proposées lors du lancement du projet DIADEMS. L'IRIT était le responsable des tâches scientifiques suivantes : détection des zones d'intérêt (démarrage des sessions d'enregistrement), détection de la musique, détection du chant qui comporte un sous-prétraitement de séparation Monophonie / Polyphonie notée « MonoPoly », segmentation en tours de chant et regroupement en chanteurs. La chaîne de traitement complète appliquée sur le corpus DIADEMS est présentée dans la Figure 6. 1.

Les modules entourés en rouge correspondent à des contributions de ma thèse dont j'ai effectué les évaluations. Les modules entourés en bleu sont évalués par les ethnomusicologues. Le module avec un contour sous forme de tirets n'a pas été encore évalué. Les trois premiers modules de cette chaîne sont considérés comme une étape de prétraitement pour le système de segmentation et regroupement en chanteurs.

[7] http://www.irit.fr/recherches/SAMOVA/DIADEMS/

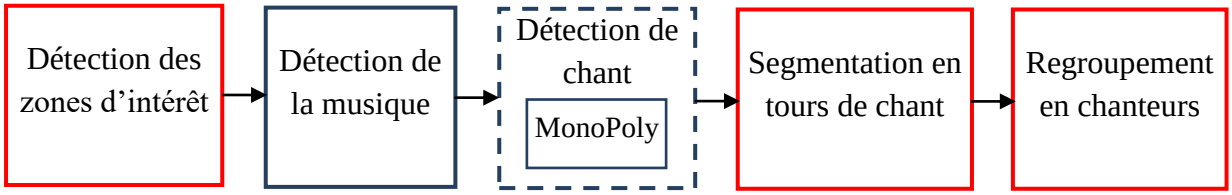

Figure 6. 1 – Architecture générale d'une chaîne de traitement appliquée sur le corpus DIADEMS.

Au cours de ce chapitre, nous décrivons, dans une première partie, le corpus DIADEMS. La deuxième partie est consacrée à la présentation des différentes étapes de prétraitement appliquées à ce corpus ainsi que les outils utilisés : détection des zones d'intérêts, détection de la musique et détection de chant. L'évaluation de notre système de segmentation en tours de chant est présentée dans la troisième partie. La quatrième partie est dédiée à l'évaluation du système de regroupement en chanteurs, ainsi qu'aux possibilités d'amélioration.

6.2. Corpus DIADEMS

Les enregistrements du projet DIADEMS viennent des archives sonores du CREM qui sont parmi les plus importantes d'Europe, en termes de qualité, de quantité et de diversité. La constitution de ces archives a commencé depuis la fin du 19ème siècle quand les premiers appareils enregistreurs ont été inventés. Les enregistrements sont réalisés par des chercheurs en musique qui se déplacent sur tous les continents afin d'alimenter et enrichir ces archives. Une petite partie de ces archives a été publiée en disque 78 tours, disque 33 tours et en CD. Une numérisation des supports analogiques est en cours. L'origine du corpus sonore du projet DIADEMS explique la grande variété de ses enregistrements en terme de contenu et de formes : musique instrumentale, voix parlée, voix chantée, voix récitée, voix psalmodiée, bruits, etc. Ces enregistrements sont très hétérogènes vu la variété des conditions et supports d'enregistrement utilisés, qui changent en avançant dans le temps. De plus, des bruits techniques liés aux supports d'enregistrement sont présents dans certains fichiers sonores de DIADEMS, ce qui nous a conduits à proposer une méthode afin de les détecter. Nous décrivons les différents types de bruits techniques rencontrés dans la section 6.3.

Dans cette partie, nous présentons le sous-corpus DIADEMS qui contient des enregistrements spécifiques à notre tâche de segmentation et regroupement en chanteurs. Ce sous-corpus a été sélectionné par les ethnomusicologues afin de pouvoir étudier et caractériser les tours de chant.

Nous avons déjà défini, au cours du Chapitre 3, le terme « **tour de chant** ». Nous avons également présenté le guide d'annotation que nous avons mis en place pour pouvoir créer la vérité terrain (références). Pour cela, nous nous limitons ici, à décrire le sous-corpus DIADEMS sélectionné pour la détection des tours de chant ainsi que quelques situations rencontrées lors de son annotation.

6.2.1. Description du sous-corpus DIADEMS, dédié aux tours de chant

La plupart des enregistrements du corpus DIADEMS choisis pour la détection des tours de chant ont été réalisés entre les années 1940 et 1980 dans plusieurs pays sub-sahariens (Congo, Gabon, Cameroun), avec une qualité acoustique variable : enregistrements en extérieur en général, présence de bruits de fond et d'évènements sonores autres que la musique. Des exemples sont accessibles en ligne, via la plateforme Telemeta [8]. Nous convertissons ces enregistrements sonores afin d'avoir les mêmes caractéristiques techniques que celles du corpus « studio » : 16 bits, 16 kHz et mono.

Les enregistrements contiennent des tours de chant soliste/chœur, des zones de voix chantée en alternance ou superposées avec des instruments ou de la parole. Ce corpus est constitué de 9 enregistrements d'une durée totale d'environ 18 minutes que nous avons divisées là aussi en un ensemble de développement (DEV) et un ensemble d'évaluation (EVAL) dans les proportions 20% et 80%. Le Tableau 6. 1 illustre la répartition du corpus spécifique pour la détection des tours de chant du projet DIADEMS. L'ensemble DEV est composé d'alternances entre 14 groupes de chanteur(s) et sert à ajuster le paramètre S_0 du système de segmentation en tours de chant. L'ensemble EVAL est utilisé pour évaluer la performance de notre système de segmentation et regroupement en chanteurs sur des enregistrements ethnomusicologiques : il comporte 41 groupes de chanteur(s).

Tableau 6. 1 – Répartition et description du corpus de détection des tours de chant du projet DIADEMS.

Ensembles	**DEV DIADEMS**	**EVAL DIADEMS**
Nombre d'enregistrements	3	6
Durée	220 secondes	850 secondes
Nombre de groupes de chanteurs	14	41

Les différentes situations de changement rencontrées dans ce corpus sont les mêmes que celles rencontrées dans le corpus « studio », qui sont détaillées dans les sections 3.2. et 3.3.2. La plupart des enregistrements DIADEMS font partie de la première situation de changement qui correspond au passage d'un groupe de chanteur(s) à un autre groupe de chanteur(s). La deuxième situation de changement, qui correspond au passage d'une zone de chant à une zone de non-chant, est aussi présente mais les zones de non-chant sont souvent des zones de parole ou de musique instrumentale, alors que pour le corpus « studio » les zones de non-chant étaient du silence.

Les enregistrements du corpus DIADEMS contiennent beaucoup de bruits de fond. Nous trouvons souvent du chant accompagné avec des instruments, du chant superposé avec de la parole en fond ou du chant accompagné avec d'autres événements sonores : frappement de

[8] http://diadems.telemeta.org/archives/DIADEMS/

mains, cloches, bruits d'animaux… La Figure 6. 2 montre le spectrogramme d'un exemple du corpus DIADEMS qui contient du chant accompagné de frappements de mains en fond en continu dans tout l'enregistrement. Les segments annotés avec l'étiquette « G1 » contiennent le chant d'un chœur et les segments annotés avec l'étiquette « G2 » contiennent le chant d'un soliste. Le trait vertical entouré en rouge représente un exemple de frappement des mains. La Figure 6. 3 présente le spectrogramme d'une région d'un fichier du corpus DIADEMS qui contient du chant accompagné de cloches en fond. Le segment annoté avec l'étiquette « G2 » correspond à un segment de non-chant. Comme nous pouvons l'apercevoir dans ce segment, il y a une forte énergie présente, engendrée par la présence des cloches. Les segments « G3 » et « G4 » contiennent le chant d'un soliste et d'un chœur respectivement, avec des cloches en fond. La région du spectrogramme entourée en rouge représente les cloches.

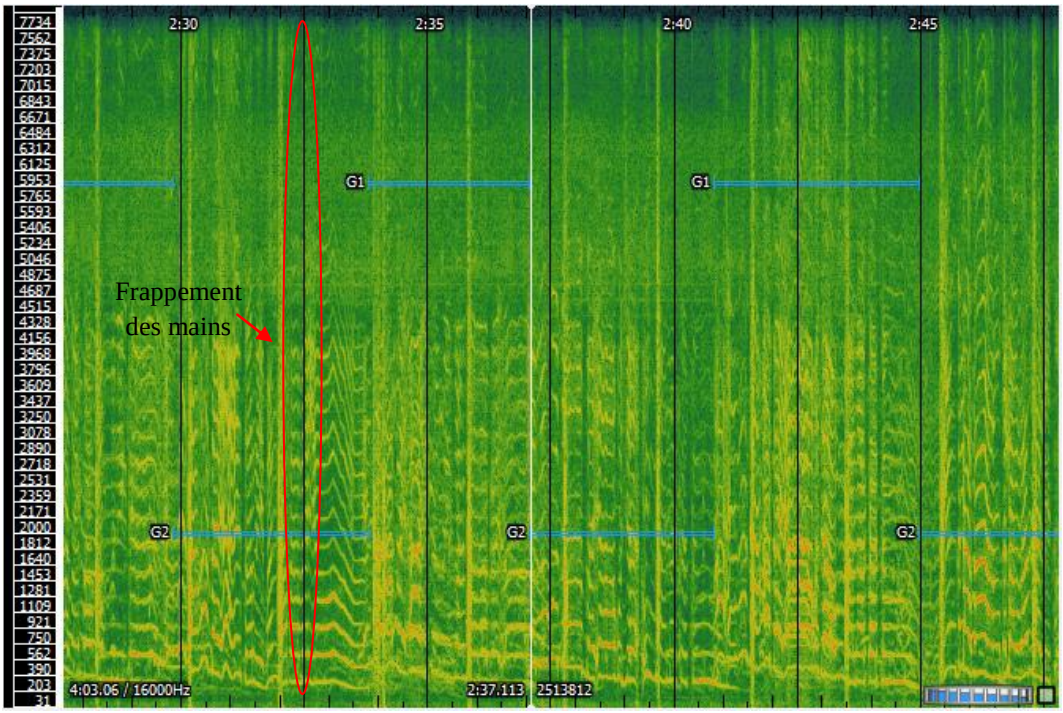

Figure 6. 2 – Spectrogramme d'un enregistrement du corpus DIADEMS contenant du chant accompagné de frappements des mains. Il s'agit d'un extrait de 20 secondes du fichier « tour_de_chant_solo_choeur_frappement_mains ».

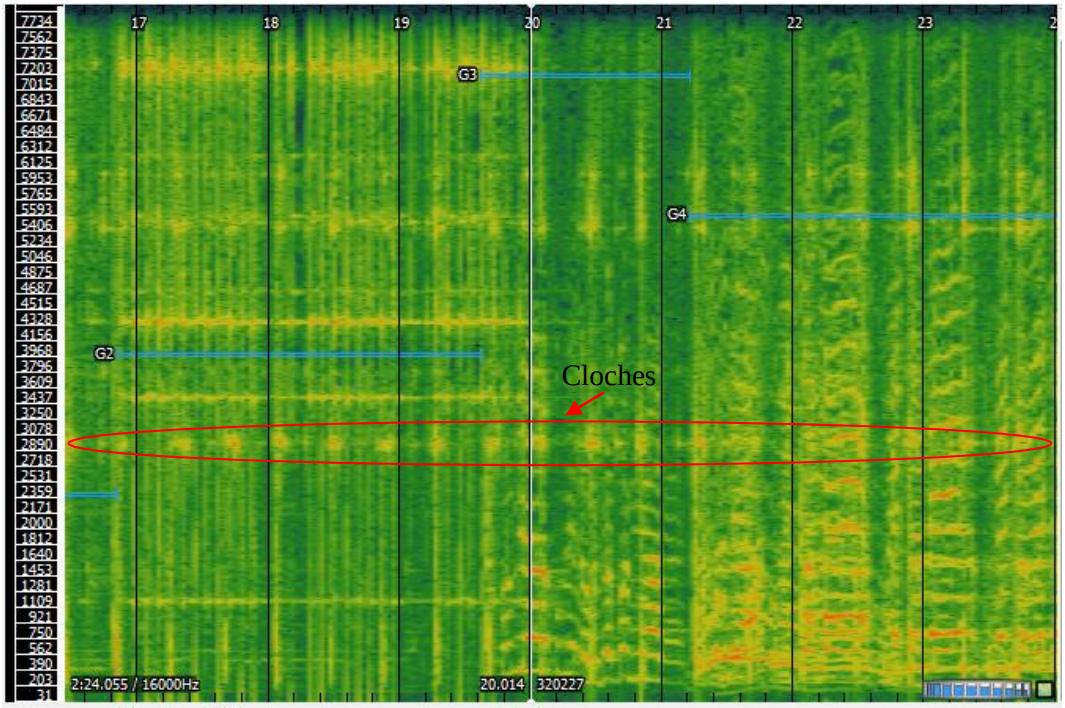

Figure 6. 3 – Spectrogramme d'un enregistrement du corpus DIADEMS contenant du chant accompagné du bruit des cloches. Il s'agit d'un extrait de 10 secondes du fichier « tours_de_chant_soliste_choeur2 ».

6.2.2. Annotation manuelle du corpus

Afin de pouvoir évaluer la performance de notre système de segmentation et regroupement en chanteurs sur les enregistrements du DIADEMS, nous avons annoté manuellement les enregistrements de ce corpus en groupes de chanteur(s), en utilisant le guide d'annotation que nous avons défini. Toutes les conventions et conditions d'annotation de ce guide sont détaillées dans la section 3.4.

Lors de l'annotation de ce corpus, nous avons rencontré plusieurs cas difficiles à annoter tels que le cas présenté dans la Figure 6. 4, qui montre le spectrogramme d'un exemple du corpus DIADEMS qui contient des alternances rapides entre un soliste (segments verts portant l'étiquette « S ») et un chœur (segments rouges portant l'étiquette « Ch »). Il était difficile de trouver les frontières modélisant les passages entre le soliste et le chœur car le soliste chante parfois avec le chœur.

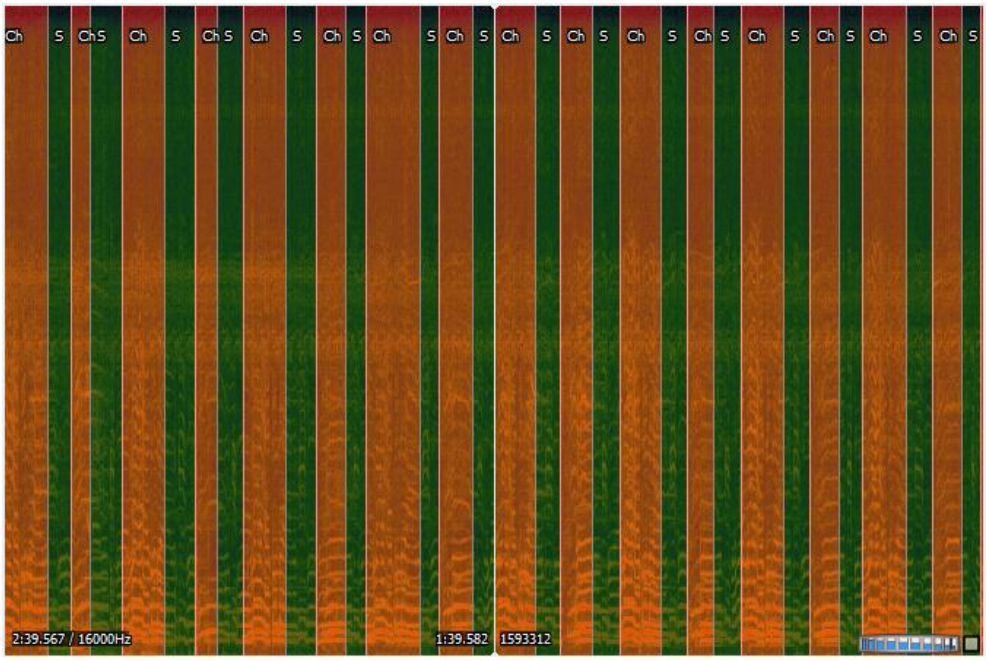

Figure 6. 4 – Spectrogramme d'un enregistrement du corpus DIADEMS difficile à annoter. Il s'agit d'un extrait de 30 secondes du fichier « altern_superp_chant_parole ».

6.3. Prétraitement appliqué au corpus DIADEMS

Dans cette section, nous décrivons les trois premiers modules de la chaîne de traitement complète appliquée sur DIADEMS qui est illustrée dans la Figure 6. 1.

6.3.1. Détection des zones d'intérêt

6.3.1.1. Les types de bruit technique

La détection des zones d'intérêt consiste à repérer, dans les enregistrements DIADEMS, les régions intéressantes pour les ethnomusicologues et ethnolinguistes. Ces zones contiennent des sons musicaux et/ou linguistiques. La plupart des enregistrements du corpus contiennent des bruits appelés « bruits techniques » qu'il convient de détecter pour les masquer et permettre de se concentrer par la suite uniquement sur les zones d'intérêt. Nous nous sommes donc attachés à mettre en œuvre des techniques pour détecter automatiquement ces bruits, qui correspondent à des défauts techniques d'enregistrement et/ou des bruits de support d'enregistrement. Comme les enregistrements DIADEMS sont très hétérogènes (enregistrements en extérieur, sur différents types de support, de 1900 à nos jours…), beaucoup de bruits techniques peuvent être présents. Ces bruits ont été classés par les acousticiens du projet selon les trois catégories suivantes.

- Début/fin d'une session de prise de son : dans cette catégorie, nous trouvons :
    - le « cloc » ou le « plop » correspond à un son bref et très basse fréquence. Il est caractéristique du démarrage et arrêt de magnétophone à bande du type Nagra ou

du démarrage et arrêt du cylindre. Ce phénomène est illustré entre les deux traits violets dans la Figure 6. 5.

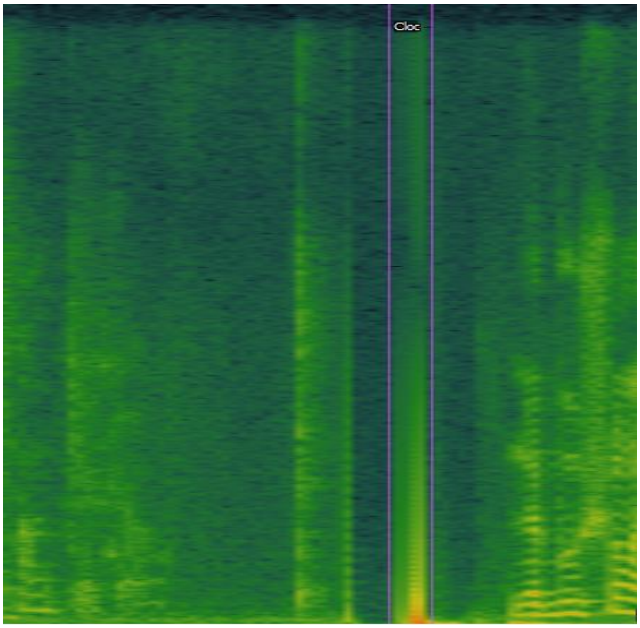

Figure 6. 5 – Illustration du bruit technique « cloc » sur un exemple des enregistrements DIADEMS.

- le « 1/x » caractérise le démarrage et arrêt de magnétophone à bande du type Nagra. Il s'agit d'une baisse rapide de toutes les fréquences. L'IRIT a appelé ce type de bruit « 1/x » parce qu'il désigne la forme des courbes de fréquence visualisée sur un spectrogramme, ce phénomène est montré dans la Figure 6. 10.

- le « blanc » signifie aucun signal sonore.

- Rupture ou discontinuité du signal : cette catégorie comporte également trois types :

  - le « drop out » (analogique ou numérique) est une absence du signal utile à la lecture à cause de la détérioration du support (morceaux de couche magnétique arrachée du ruban ou cassette DAT détériorée). Il est illustré dans la Figure 6. 6.

  - le « glitch » est un bruit parasite sur les DAT numériques, i.e. un grésillement caractéristique des distorsions numériques liées probablement à un dépassement des capacités des correcteurs d'erreur. Il est généralement accompagné d'un « drop out » et il est montré aussi dans la Figure 6. 6.

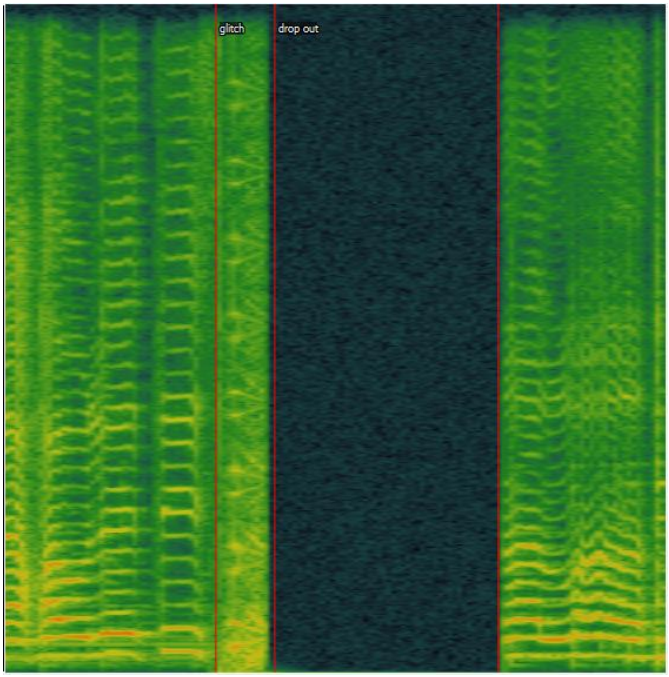

Figure 6. 6 – Illustration du bruit technique « glitch » et du phénomène « drop out » sur un enregistrement du corpus DIADEMS.

- le « Crac » est un son typique du disque vinyle lors de la lecture qui permet de caractériser les supports mécaniques (33t, 45t). La Figure 6. 7 illustre ce type de bruit, qui est isolé entre les deux traits rouges.

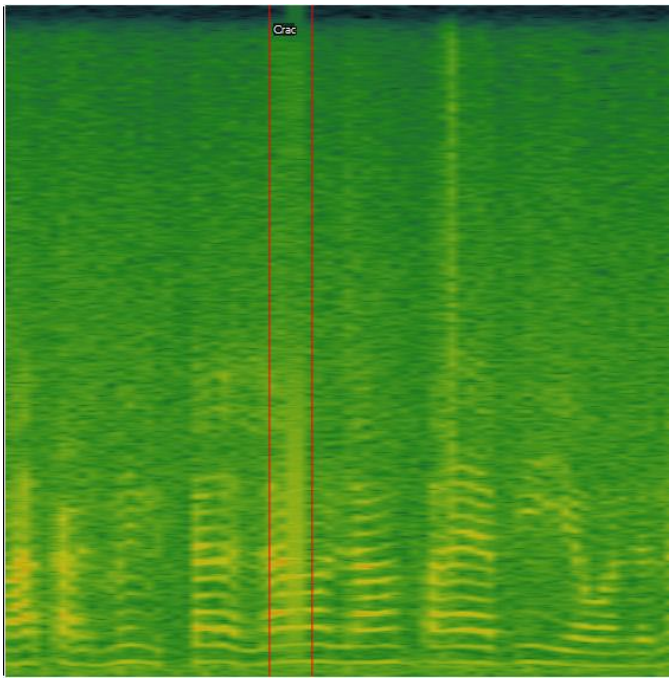

Figure 6. 7 – Illustration du bruit technique « crac » sur un enregistrement du corpus DIADEMS.

- Défauts techniques pendant la prise de son : dans cette catégorie, nous trouvons :
    - la « saturation », une distorsion produite par l'amplitude excessive du signal sonore ; elle est illustrée dans la Figure 6. 8.

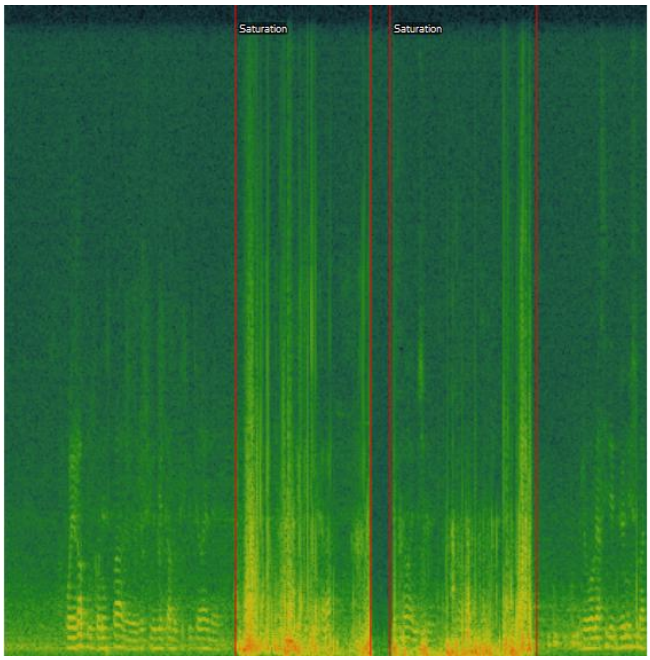

Figure 6. 8 – Illustration du phénomène de « saturation » sur un enregistrement du corpus DIADEMS.

- la « chute » ou la « montée » brusque du niveau sonore d'enregistrement. La Figure 6.9 montre un exemple de phénomène d'une montée brusque.

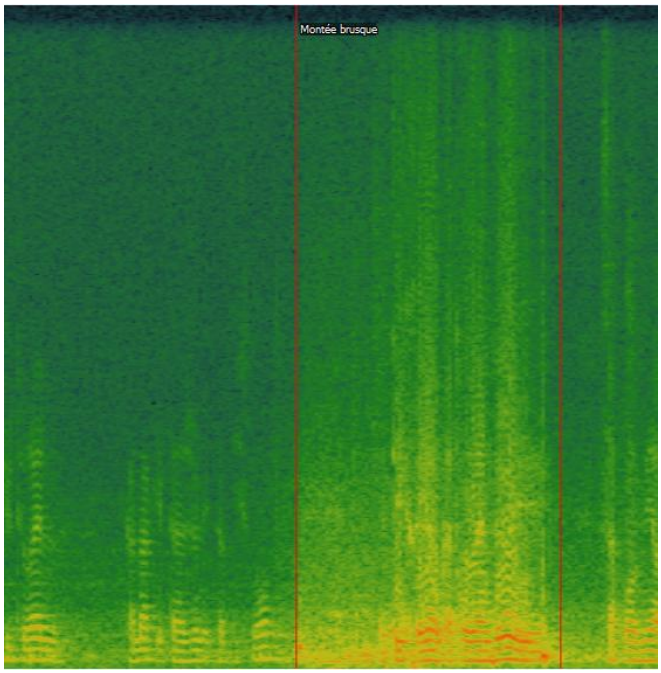

Figure 6. 9 – Illustration d'un exemple d'une montée brusque sur un enregistrement du corpus DIADEMS.

- les bruits de « micro » peuvent être un choc des microphones, un son du vent dans le micro ou un frottement des câbles du micro ou de la main tenant le micro.

6.3.1.2. Deux algorithmes de détection des bruits de type « 1/x »

L'IRIT a été en charge de la détection des bruits indiquant le début et la fin d'une session d'enregistrement, i.e. la détection du bruit dit « 1/x ». La Figure 6. 10 montre le phénomène de décroissance en « 1/x », qui est isolé entre les deux frontières rouges, sur chaque exemple.

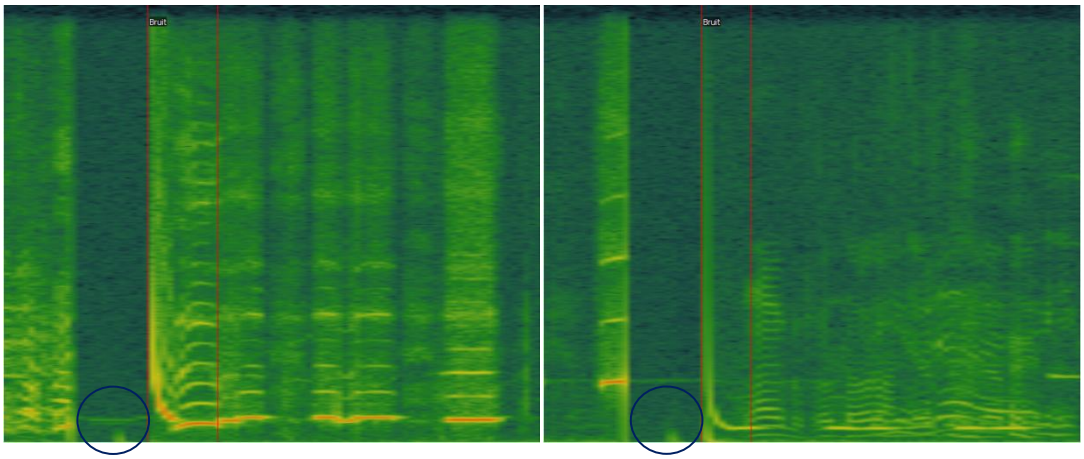

Figure 6. 10 – Illustration du bruit technique « 1/x » sur deux exemples des enregistrements du corpus DIADEMS.

a. Repérage de la décroissance « 1/x »

Nous avons commencé par développer une méthode de détection de ce bruit en essayant de repérer la décroissance « 1/x ». La Figure 6. 11 montre les étapes de cette méthode. La première étape consiste à extraire les paramètres énergie spectrale et centroïde spectral. La deuxième détecte les zones d'arrêt du signal en utilisant un seuil sur l'énergie spectrale égal à 0,5. La troisième extrait une partie de la zone suivant l'arrêt du signal, de durée 0,42 s qui est la durée maximale du phénomène « 1/x » (durée du segment orange qui porte l'étiquette « phénomène 1/x » dans la Figure 6. 12). Ensuite, un lissage médian de la courbe du centroïde spectral (courbe en rouge dans la Figure 6. 12) pour la partie extraite est réalisé et suivi d'un calcul du gradient pour calculer le taux de décroissance (TD). Si le TD est supérieur à 50%, nous effectuons une extraction des pics locaux et un calcul des hauteurs ΔC1, ΔC2 et ΔC3 entre les différents pics est réalisé (cf. Figure 6. 12). ΔC1 est la différence entre le premier pic, qui est le début de la décroissance, et le troisième pic. ΔC2 représente l'hauteur entre un pic local et le pic minimal qui le précède. ΔC3 est la différence entre le pic de début et de fin de la fenêtre de décroissance. Afin de décider s'il s'agit du phénomène « 1/x » ou non, un seuillage est effectué : le ΔC1 doit avoir une valeur strictement positive, ΔC2 et ΔC3 doivent être supérieurs à 110 Hz et 300 Hz respectivement. Ces deux seuils ont été fixés sur quelques exemples du phénomène « 1/x ».

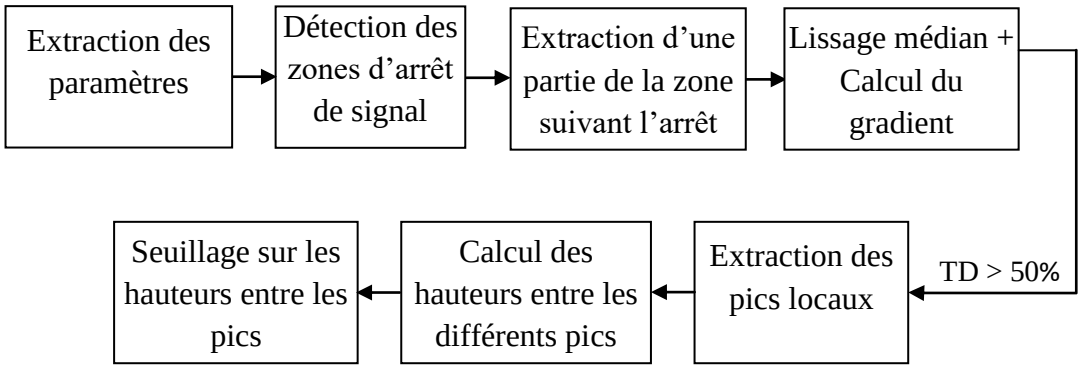

Figure 6. 11 – Illustration de la méthode de détection du phénomène « 1/x ».

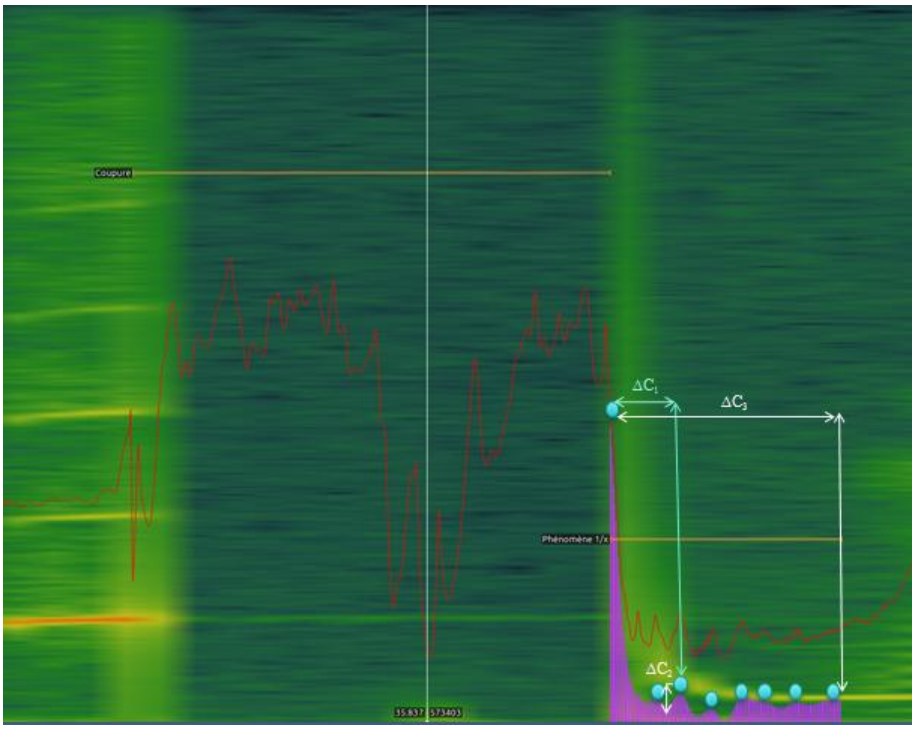

Figure 6. 12 – Présentation du déroulement de l'algorithme de détection du phénomène « 1/x ». La courbe en rouge représente le centroïde spectral et la courbe violète représente le centroïde spectral après lissage médian sur la partie de décroissance en « 1/x ».

Nous avons testé cette méthode sur 12 exemples du phénomène « 1/x » du corpus DIADEMS. Neuf exemples sur douze ont été correctement détectés (cf. Tableau 6. 2).

b. Empreinte du démarrage de bande

Pour aller plus loin, nous nous sommes basés sur le comportement spécifique de l'énergie qui précède le phénomène « 1/x » et qui est entouré en bleu sur chaque exemple de la Figure 6. 10. Ce comportement est caractéristique du démarrage du magnétophone à bande et donc du début d'enregistrement. Nous avons ainsi développé une méthode de détection de ce comportement pour repérer les arrêts / démarrages de bande. Elle est illustrée dans la

Figure 6. 13. Cette méthode consiste, tout d'abord, à extraire l'énergie du signal, ensuite, à apprendre, à partir de quelques exemples de ce comportement spécifique (données d'apprentissage), un modèle caractérisant le comportement de l'énergie lorsque le magnétophone démarre et commence à enregistrer. Ce modèle est une courbe d'énergie réalisée après alignement de tous les exemples d'apprentissage et après extraction de toutes les courbes d'énergie pour chaque exemple. Puis, la moyenne point à point de toutes les courbes d'énergie est calculée afin d'avoir un modèle final qui sera considéré comme une empreinte audio qui va être recherchée, par la suite, dans le signal audio par notre algorithme. La recherche de cette empreinte est effectuée en repérant, tout d'abord, toutes les zones de silence dans le signal en utilisant un seuil sur l'énergie égal à 0,002, et en considérant par la suite celles qui sont de durée supérieure à 0,2 s. Puis nous traçons la courbe de différence moyenne d'énergie qui est obtenue en faisant glisser notre empreinte dans toute la zone de silence courante et en calculant la différence d'énergie point à point. La moyenne de toutes les différences point à point correspond à un point de la courbe. Ensuite, nous cherchons le minimum de cette courbe et si sa valeur est inférieure à un certain seuil, égal à 0,05, alors il s'agit d'un arrêt / démarrage de bande.

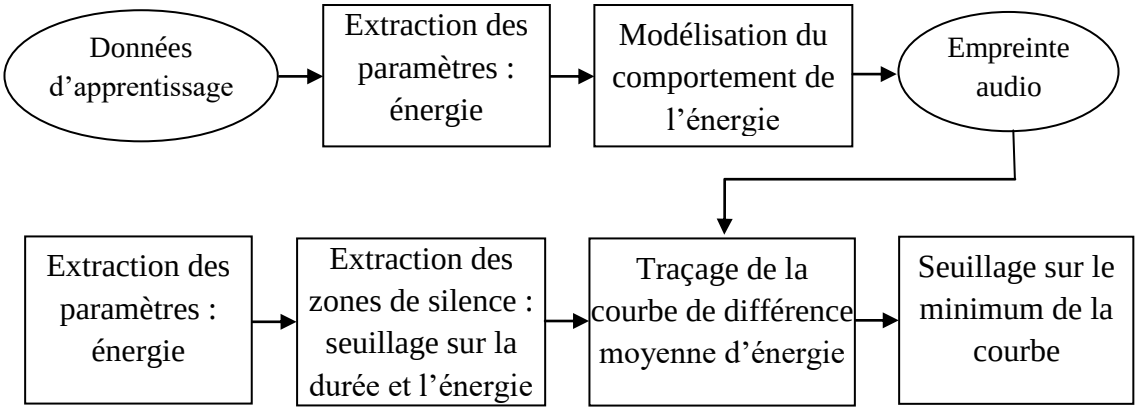

Figure 6. 13 – Illustration de la méthode de détection du comportement spécifique de l'énergie qui précède le phénomène « 1/x ».

Tableau 6. 2 – Résultats de la détection de démarrages / arrêts de bande.

Méthodes	**Nombre d'items (vérité terrain)**	**Nombre d'items correctement détectés**
Détection de « 1/x »	12	9
Détection du comportement spécifique de l'énergie	12	12
	11	9
	23 (total)	21

Avec cette méthode, nous avons réussi à détecter tous les arrêts / démarrage de bandes dans les 12 exemples sur lesquels nous avons testé la méthode de détection du phénomène

« 1/x » (cf. Tableau 6. 2). Nous l'avons aussi appliquée sur 11 autres exemples pour lesquels il y avait 9 détectés parmi les 11, ce qui implique un nombre total de 21 items correctement détectés parmi les 23 (total des exemples testés). Cela prouve que cette méthode est plus performante. Le fait que la détection ne fonctionne pas pour deux exemples est lié à la présence d'un bruit de fond trop important.

6.3.2. Détection de la musique

Les outils utilisés pour la détection de la musique (Pinquier, et al., 2002) s'appuient sur une segmentation automatique du signal en des zones dites « homogènes », issue de l'algorithme « Divergence Forward-Backward » (André-Obrecht, 1988). L'algorithme est fondé sur une étude statistique du signal dans le domaine temporel en faisant l'hypothèse que le signal est décrit par une suite de zones quasi stationnaires, chacune est caractérisée par un modèle statistique autorégressif. La méthode consiste à détecter les changements de ces modèles au travers des erreurs de prédiction calculées sur deux fenêtres d'analyse successives. La distance entre les deux modèles est obtenue à partir de l'entropie mutuelle des deux lois conditionnelles correspondantes. A partir de cette segmentation en zones homogènes, deux paramètres sont calculés afin de décider s'il y a de la musique ou non :

- Le premier consiste à calculer, sur une seconde du signal, la longueur moyenne des segments significatifs (segments ayant une taille suffisante). La détection de la musique est basée sur le fait que, pour la musique, les segments ont tendance à être beaucoup plus longs qu'en parole car il y a des phases plus longues (durées des notes plus longues que celles des phonèmes en général).

- Le second extrait le nombre de segments par seconde. La détection de la musique par cet outil est basée sur le fait qu'en musique ces segments ont tendance à être beaucoup moins nombreux.

Les deux paramètres sont fortement corrélés mais peuvent donner des résultats assez différents. Ils ont été testés sur un corpus réalisé à partir des enregistrements de RFI (Radio France Internationale) dans le cadre de la thèse de Julien Pinquier (Pinquier, 2004). Le taux de classification correcte obtenu sur ce corpus est de 86,4% avec le paramètre « nombre de segments » et de 78,1% avec le paramètre « durée des segments ». Ces deux approches ont été évaluées pour le corpus DIADEMS par les ethnomusicologues en utilisant comme métrique d'évaluation un taux d'efficacité. Le taux d'efficacité correspond au rapport entre la durée correctement détectée dans un item donné et la durée totale de l'item. Les résultats de l'analyse automatique sont confrontés seconde par seconde avec la catégorisation obtenue par le travail d'écoute effectué par les ethnomusicologues. Une addition des durées des portions correctement détectées est effectuée afin d'avoir la durée totale correctement détectée par l'outil. Le taux d'efficacité a été calculé pour 25 exemples testés par ces deux paramètres. Les résultats sont illustrés dans le Tableau 6. 3. Un taux d'efficacité de 68% est obtenu en utilisant le paramètre « nombre de segments » et de 53% avec le paramètre « durée des segments ». Nous pouvons noter que le paramètre « nombre de segments » est plus efficace que « durées des segments » pour les deux corpus. Les deux paramètres détectent correctement la voix

chantée et les instruments de musique comme *musique*. Ils détectent également la voix parlée comme *non musique*.

Les erreurs de détection concernent les items présentant de la voix chantée simultanée à des instruments de musique de type percussif. Ces derniers sont d'intensité plus forte et ont des segments plus courts. Les deux approches détectent des évènements sonores comme la voix racontée ou le silence comme *musique*, les segments étant plus longs et donc moins nombreux aussi. Ils détectent aussi le chant ou un instrument de musique comme *non musique* lorsque certaines syllabes et trémolos sont très serrés.

Tableau 6. 3 – Efficacité des deux paramètres de détection de la musique.

Paramètres	**Taux d'efficacité**
« Nombre de segments »	68%
« Durée des segments »	53%

6.3.3. Détection du chant

Après avoir détecté les régions contenant de la musique, un module de détection du chant peut être appliqué afin de détecter les zones de chant (celles-ci serviront d'entrée à notre système de segmentation et regroupement en chanteurs). La méthode utilisée est décrite dans la section 2.3.1 (Lachambre, et al., 2009a). L'étape de détection du chant est précédée d'une étape de séparation Monophonie / Polyphonie (« MonoPoly »).

6.3.3.1. Séparation monophonie / polyphonie

Le module de détection monophonie / polyphonie utilise l'estimateur de fréquence fondamentale YIN ainsi que les moyennes et variances de l'harmonicité calculées sur une seconde du signal pour savoir si celle-ci contient plutôt une ou plusieurs sources. Ce module a été testé sur un corpus studio fait « maison » réalisé dans le cadre de la thèse d'Hélène Lachambre (Lachambre, 2009b) que nous appelons ici corpus « Lachambre ». Il contient des sons monophoniques (instrument solo, chanteur solo) et des sons polyphoniques (plusieurs instruments, plusieurs chanteurs, instruments et chanteurs). Le résultat de ce module sur ce corpus était très satisfaisant : un faible taux d'erreurs de 6,3% a été obtenu. L'évaluation de ce module sur le corpus DIADEMS a été faite par les ethnomusicologues en utilisant la même métrique utilisée pour la détection de la musique qui est le taux d'efficacité. Un taux moyen d'efficacité de 90% a été obtenu sur 12 items testés du corpus DIADEMS.

Les items évalués contiennent de la musique instrumentale avec un seul instrument, musique instrumentale avec plusieurs instruments, contenu vocal (chant, parole) et de la musique voco-instrumentale (chant accompagné des instruments). De très bons résultats sont obtenus pour les items contenant de la musique instrumentale. La production sonore de chacun des instruments, pour la musique instrumentale avec un seul instrument, est identifiée comme de la *monophonie*. La détection de la *polyphonie* résultant du jeu de plusieurs instruments donne également de très bons résultats. Dans le cas d'un ensemble d'instruments

appartenant à la même famille instrumentale et jouant des parties musicales différentes, la production sonore est détectée comme de la *polyphonie*. Les courts segments détectés comme de la *monophonie* correspondent à des sons prolongés, joués à l'unisson ou à intervalle de tierce mineure. En ce qui concerne la production vocale chantée, il faut distinguer le chant solo, d'une part, et le chœur, d'autre part. Selon les exemples évalués, le chant solo est détecté correctement comme de la *monophonie* avec un taux efficacité compris entre 50% et 100%. Les résultats concernant les parties du chœur sont équivalents à ceux du chant solo. Lorsque les lignes mélodiques du chœur sont différentes, les interventions de ce dernier sont détectées comme de la *polyphonie* avec un taux efficacité compris entre 50% et 95%. Quand le chœur est à l'unisson, il est bien détecté comme de la *monophonie*.

Ensuite, les interventions voco-instrumentales sont très bien détectées comme de la *polyphonie* (efficacité entre 80% et 100%). Le silence est détecté, dans la majorité des cas, comme de la *polyphonie*.

6.3.3.2. Détection de chant

Le détecteur de chant est basé sur la détection du vibrato en tenant compte du contexte monophonique et polyphonique. Ce détecteur donne un taux d'erreur de 25% sur le corpus « Lachambre ». A l'heure où j'ai écrit la thèse, l'évaluation de ce détecteur sur le corpus DIADEMS n'a pas été effectuée. Pour cela, nous avons préféré utiliser un corpus spécifique, qui ne contient que du chant, pour la partie suivante de segmentation et regroupement en chanteurs.

6.4. Segmentation en tours de chant

Après avoir validé nos systèmes de segmentation en tours de chant et regroupement en chanteurs sur le corpus « studio » (cf. Chapitre 4 et Chapitre 5), nous les appliquons ici sur le corpus DIADEMS.

6.4.1. Application de notre système de segmentation en tours de chant

Nous rappelons à partir de la Figure 6. 14 l'architecture générale de notre système de segmentation en tours de chant, son architecture étant détaillée dans le Chapitre 4.

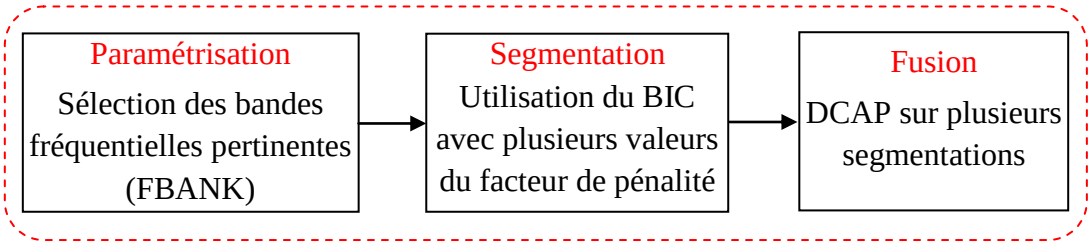

Figure 6. 14 – Rappel de l'architecture générale de notre système de segmentation en tours de chant.

L'application de ce système sur les enregistrements DIADEMS avec les mêmes configurations que pour le corpus « studio » implique, tout d'abord, une première étape de sélection des bandes fréquentielles pertinentes qui consiste à choisir parmi les douze premières bandes celles qui possèdent une variance élevée en obéissant aux règles de la stratégie de sélection des bandes que nous avons mise en place et qui est détaillée dans la section 4.5.4. Ensuite, plusieurs segmentations sont réalisées en faisant varier la valeur du coefficient de pénalité λ sur l'intervalle $[2{,}0 \;\; 10{,}0]$ avec un pas de 0,05, donc 161 systèmes sont obtenus. Puis, la méthode DCAP est appliquée sur les 161 sorties résultantes de l'étape de segmentation afin d'effectuer un vote majoritaire. Cette dernière étape nécessite l'ajustement du seuil S_0 de la méthode DCAP. Comme pour le corpus « studio », nous ajustons ce seuil sur un ensemble de développement qui est dans ce cas le DEV du corpus DIADEMS, décrit dans la section 6.2. En ajustant ce seuil sur le DEV, nous trouvons une valeur de S_0 égale à 3. Nous appliquons sur l'ensemble d'évaluation (EVAL) du corpus DIADEMS la valeur de ce seuil pour la méthode DCAP.

Afin de connaître la performance maximale que nous pouvions atteindre sur le corpus DIADEMS avec notre système de segmentation, nous avons extrait la performance du système « *oracle* » de segmentation par BIC en considérant pour chaque fichier la meilleure valeur de λ (celle permettant d'obtenir la F-mesure la plus élevée).

6.4.2. Résultats de la segmentation en tours de chant sur DIADEMS

Les résultats du système « *oracle* » ainsi que de notre système de segmentation en tours de chant « complet » avec la méthode DCAP sur le DEV et l'EVAL du corpus DIADEMS sont rassemblés dans le Tableau 6. 4. Ces performances sont obtenues en utilisant une tolérance de 0,5 secondes sur les frontières des segments.

Avec notre système complet de segmentation en tours de chant (FBANK sélectionnés, segmentation par BIC, DCAP), nous obtenons une performance globale de 66,4% et 61,4% en termes de F-mesure sur le DEV et l'EVAL, respectivement. Naturellement cette performance reste plus faible que la performance du système « *oracle* » : il existe encore une marge de gain possible de 5,6% sur EVAL, si on considère la F-mesure obtenue avec le système « *oracle* » comme la limite supérieure à atteindre.

Tableau 6. 4 – Résultats des systèmes de segmentation en tours de chant « oracle » et DCAP sur le DEV et l'EVAL du corpus DIADEMS.

Systèmes	**DEV DIADEMS**			**EVAL DIADEMS**		
	Précision	**Rappel**	**F-mesure**	**Précision**	**Rappel**	**F-mesure**
« *oracle* »	76,4	62,0	68,4	60,3	73,0	66,0
DCAP – $S_0 = 3$	69,8	63,3	66,4	52,4	74,2	61,4

Nous pouvons noter que les performances sont inférieures à celles obtenues sur le corpus « studio ». Cela peut être dû à la difficulté accrue causée par la présence d'instruments et par la qualité sonore variable des enregistrements DIADEMS. En effet, les enregistrements du corpus « studio » contenaient seulement du chant, alors que les enregistrements du corpus DIADEMS sont très variés. Ainsi, certains enregistrements DIADEMS, contenant seulement du chant, montrent une performance de 80% alors que d'autres ne dépassent pas 40%. Les erreurs sur ces enregistrements sont principalement des fausses alarmes : leur écoute révèle beaucoup de superpositions de chanteurs, d'alternances très rapides entre solistes et chœur, la présence d'instruments percussifs et du bruit de fond (principalement de la parole et des cris). En outre, ces enregistrements se sont avérés être les plus difficiles à annoter manuellement : dans certains cas d'alternances rapides de chanteurs, il n'était pas évident de décider s'il fallait réellement insérer une frontière de segment ou non.

6.5. Regroupement en chanteurs

Lors du chapitre 5, j'ai proposé un premier système de regroupement en chanteurs (RDCAP) et lorsque j'ai ajouté deux passes supplémentaires à ce système, j'ai obtenu un deuxième système de regroupement (RDCAP+VCE) qui a permis d'améliorer le résultat de regroupement sur le corpus « studio ». Pour cela, nous appliquons, au cours de cette section, les deux systèmes sur le corpus DIADEMS. Des expériences complémentaires sont proposées pour améliorer la performance de la tâche segmentation et regroupement en chanteurs sur le corpus DIADEMS.

6.5.1. Application du système de regroupement en chanteurs RDCAP

6.5.1.1. Système de regroupement en chanteurs RDCAP

L'architecture générale du système de regroupement en chanteurs RDCAP est rappelée dans la Figure 6. 15. Pour plus de détails, voir le Chapitre 5.

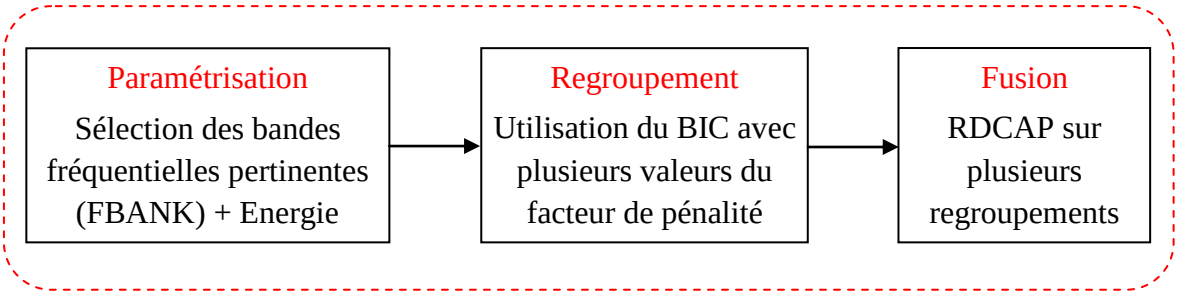

Figure 6. 15 – Rappel de l'architecture générale de notre système RDCAP de regroupement en chanteurs.

L'application de ce système sur les enregistrements DIADEMS implique une première passe de paramétrisation qui est effectuée pour extraire les FBANK sélectionnés et l'énergie. Ensuite, le regroupement par BIC est appliqué sur la segmentation automatique résultante de notre système de segmentation en tours de chant. Plusieurs systèmes de regroupement sont réalisés en faisant varier la valeur du facteur de pénalité λ dans le même intervalle choisi pour

le corpus « studio » pour le regroupement, qui est [5,0 12,0] avec un pas de 0,5. Ainsi, 15 systèmes de regroupement sont obtenus. Puis, la méthode RDCAP est appliquée : toutes les sorties de ces systèmes sont concaténées et un vote majoritaire est effectué : chaque segment est classé dans la classe choisie par la majorité des systèmes.

Afin d'estimer le taux d'erreur le plus faible que nous pouvons avoir en utilisant ce système de regroupement sur les enregistrements DIADEMS, nous avons calculé le DER du système « *oracle* » pour le regroupement. Ce système est obtenu en choisissant une valeur optimale du facteur de pénalité λ pour chaque enregistrement lors du regroupement avec le BIC (ceci correspond au taux d'erreur le plus faible en termes de DER).

6.5.1.2. Résultats du système RDCAP sur DIADEMS

Les résultats du système « *oracle* » du regroupement et du système de regroupement RDCAP sur le DEV et l'EVAL du corpus DIADEMS sont rassemblés dans le Tableau 6. 5. Les résultats sont obtenus avec une tolérance de 0,5 seconde sur les frontières. En appliquant le système de base du regroupement par BIC (sans RDCAP) et en fixant la valeur du facteur de pénalité λ dans l'intervalle [5,0 12,0] sur le DEV, nous avons obtenu un taux d'erreurs faible de 26,8% (cf. Tableau 6. 5) sur le DEV pour une valeur de λ égale à 11,5. Pour cette valeur de λ, nous trouvons un DER de 44,3% sur l'EVAL. Ce système est bien adapté au DEV, mais il généralise moins bien sur l'EVAL.

La performance sur l'EVAL est de 30,9% avec le système « *oracle* » et de 41,3% avec le système RDCAP. Il existe encore une marge importante de réduction du taux d'erreurs (environ 10%). Avec notre système RDCAP, nous gagnions 3% sur l'EVAL du corpus DIADEMS par rapport au système de base. Cela prouve l'utilité de notre méthode de fusion RDCAP.

Tableau 6. 5 – Résultats du système « oracle », système de base et du système RDCAP sur le DEV et l'EVAL du corpus DIADEMS.

Systèmes regroupement	**DEV DIADEMS DER**	**EVAL DIADEMS DER**
Système « *oracle* »	23,7%	30,9%
Système de base – λ=11,5	26,8%	44,3%
Système RDCAP	33,0%	41,3%

Néanmoins, le DER sur le corpus DIADEMS est plus élevé que celui trouvé sur le corpus « studio ». Cela est vraisemblablement dû au fait que pour certains enregistrements DIADEMS, il y a des accompagnements du chant comme des frappements de mains et des instruments percussifs (cloches, tambour…) (cf. Figure 6. 2 et Figure 6. 3) qui perturbent le processus du regroupement et qui génèrent un phénomène de sur-regroupement. Dans ces enregistrements, tous les segments qui contiennent de l'accompagnement avec un évènement

sonore autre que le chant, sont regroupés dans une seule classe bien qu'il n'y ait pas le même groupe de chanteurs dans tous les segments.

Un enregistrement DIADEMS contenant des frappements de mains (cf. Figure 6. 1) est composé d'alternances de chant entre 13 groupes de chanteur(s). Avec notre système, nous ne trouvons que 2 classes dont une correspond à toutes les zones de chant accompagné des frappements de mains.

Un autre enregistrement contenant des sons des cloches (cf. Figure 6. 3) comporte des alternances de chant entre 10 groupes de chanteurs et une classe pour les segments ne contenant que les cloches (sachant que les cloches sont présentes durant tout le fichier). Avec notre système, nous n'obtenons que 2 classes : une classe contient tous les segments de chant de différents groupes de chanteur(s) accompagné des cloches et une classe pour tous les segments contenant seulement les sons des cloches (il n'y a pas de chant).

Dans d'autres enregistrements, nous avons remarqué des confusions entre des groupes de chanteurs. La raison principale est que les deux groupes possèdent des chanteurs en commun. Notons que ces enregistrements se sont avérés être les plus difficiles à traiter lors de la vérité terrain (regroupement manuel). En effet, il n'était pas évident de mettre la même étiquette ou non pour deux groupes de chanteurs possédant des chanteurs en commun.

6.5.2. Application du système de regroupement en chanteurs RDCAP+VCE

6.5.2.1. Système de regroupement en chanteurs RDCAP+VCE

L'architecture de notre système de regroupement en chanteurs RDCAP+VCE est rappelée dans la Figure 6. 16.

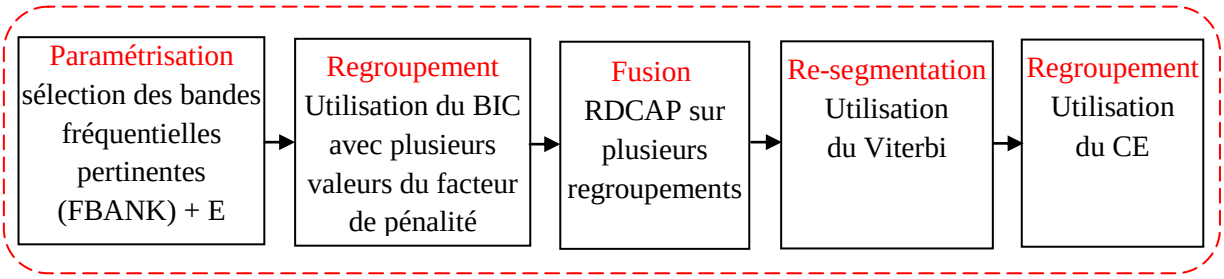

Figure 6. 16 – Rappel de l'architecture générale de notre système RDCAP+VCE de regroupement en chanteurs.

L'application de ce système sur les enregistrements DIADEMS implique l'application du système RDCAP proposé pour le regroupement, puis l'étape de re-segmentation par Viterbi et l'étape de regroupement par Entropie Croisée (CE). La dernière étape nécessite l'ajustement du seuil δ du critère CE. Pour cela, nous utilisons le DEV du corpus DIADEMS pour déterminer sa valeur en la faisant varier dans l'intervalle $[0{,}5 \;\; 2{,}5]$ avec un pas de 0,5.

Afin de comparer ce système avec les systèmes du regroupement en locuteurs, nous avons testé le système du regroupement du LIUM sur les enregistrements des tours de chant de DIADEMS. Nous avons ajusté les valeurs du couple (λ, δ) sur le DEV. Nous faisons varier la valeur du λ sur l'intervalle $[5{,}0 \;\; 12{,}0]$ avec un pas de 0,5, comme avec la méthode RDCAP.

6.5.2.2. Résultats du système RDCAP+VCE sur DIADEMS

Les résultats trouvés sur le DEV et l'EVAL avec le système du LIUM et notre système RDCAP+VCE sont illustrés dans le Tableau 6. 6. Ces résultats sont obtenus avec une tolérance de 0,5 secondes sur les frontières. Concernant le système du LIUM, le couple (λ, δ) qui nous permet d'avoir le taux d'erreurs le moins élevé en termes de DER, est $(5{,}5; 1{,}5)$. La valeur de δ qui nous donne le DER le moins élevé avec notre système RDCAP+VCE sur le DEV est égale à 1,5. Les DER pour les deux systèmes sont similaires et élevés : environ 48% sur l'EVAL.

Tableau 6. 6 – Résultats du système du regroupement du LIUM et de notre système RDCAP+VCE sur le DEV et l'EVAL du corpus DIADEMS.

Systèmes regroupement	**DEV DIADEMS DER**	**EVAL DIADEMS DER**
Système LIUM	26,1%	48,3%
Système RDCAP+VCE	26,4%	48,4%

Nous pouvons noter que ces taux d'erreurs sont plus élevés de 7% par rapport au taux obtenu avec le premier système. Cela semble prouver que notre système RDCAP est plus robuste. Cette augmentation d'erreurs est engendrée principalement par un phénomène de sur-regroupement qui apparaît déjà depuis la fin de l'étape de fusion par RDCAP, et avec les passes supplémentaires de regroupement, ce phénomène est renforcé.

6.5.3. Expériences complémentaires pour le regroupement en chanteurs sur DIADEMS

Nous avons essayé de proposer une méthode générique de segmentation et regroupement en chanteurs pour s'affranchir des conditions d'enregistrements (studio, terrain). Cependant, les performances sur les enregistrements du projet DIADEMS sont plus faibles que celles obtenues sur le corpus « studio ».

Afin de mieux répondre aux besoins des ethnomusicologues, nous avons réalisé quelques expériences complémentaires avec le système RDCAP. Notre système de regroupement étant principalement fondé sur le critère BIC, nous avons essayé de jouer sur la valeur du paramètre de pénalité λ, en testant d'autres intervalles que celui qui a été déterminé sur le corpus « studio ». En effet, si l'intervalle n'est pas bien choisi, la plupart des sorties du regroupement seront obtenus avec des valeurs de λ qui donnent des taux d'erreurs élevés, ce qui conduit à fausser le vote. Pour cela, nous avons essayé beaucoup d'intervalles sur le DEV de

DIADEMS : [2,0 5,0], [2,0 6,0], [0,5 5,5], [0,5 9,0], [0,5 12,0], etc. Nous avons ainsi sélectionné l'intervalle [0,5 9,0] avec un pas de 0,5. Les résultats obtenus avec le système « *oracle* » et notre système de regroupement RDCAP sur cet intervalle avec une tolérance de 0,5 secondes sont illustrés dans le Tableau 6. 7.

En appliquant notre système de regroupement en chanteurs RDCAP sur l'intervalle [0,5 9,0], nous obtenons un DER de 36,4% sur l'EVAL : soit une baisse de 5% du taux d'erreur par rapport au DER trouvé sur l'intervalle [5,0 12,0] (cf. Tableau 6. 5). Néanmoins, il existe encore une marge d'erreurs de 7% à combler sur l'EVAL pour arriver au même taux d'erreurs que le système « *oracle* » (29,4%).

Tableau 6. 7 – Résultats du système « oracle » et de notre système RDCAP sur le DEV et l'EVAL du corpus DIADEMS avec l'intervalle [0,5 9,0].

Systèmes regroupement	**DEV DIADEMS DER**	**EVAL DIADEMS DER**
Système « *oracle* »	27,5%	29,4%
Système RDCAP	30,1%	36,4%

Au cours des sections précédentes, j'ai validé mes contributions en évaluant les 3 modules de détection d'arrêt / démarrage de bande, de segmentation en tours de chant et de regroupement en chanteurs de manière indépendante. Comme le détecteur de chant est en cours d'évaluation, nous n'avons pas encore le taux de performance de la chaîne de traitement complète (cf. Figure 6. 4) appliquée sur le corpus DIADEMS pour réaliser de la segmentation et regroupement en chanteurs. L'étape suivante sera d'évaluer l'ensemble de la chaîne de traitement par les ethnomusicologues.

6.6. Conclusion

Dans ce chapitre, nous avons présenté le corpus du projet DIADEMS ainsi que la chaîne de traitement mise en place pour obtenir un système de segmentation et regroupement en chanteurs automatique.

La première étape de cette chaîne consiste à repérer les zones d'intérêt en détectant les bruits techniques de démarrage de session. Pour celle-ci, nous avons développé deux méthodes. La première qui consiste à détecter le phénomène de décroissance « 1/x », permet de détecter 9 parmi 12 exemples testés. La deuxième qui utilise une empreinte pour modéliser le silence spécifique précédent le démarrage de bande, permet de détecter 21 phénomènes parmi 23 évalués, ce qui est très satisfaisant comme résultat.

La deuxième étape est une détection de musique à partir de deux paramètres (durée des segments et nombre de segments). Ils ont été évalués par les ethnomusicologues : un taux

d'efficacité de 53% avec le paramètre « durée des segments » et de 68% avec le paramètre « nombre de segments » sont obtenus.

La troisième étape de notre système est une détection du chant. Celle-ci débute par une séparation monophonie / polyphonie très performante. L'application de ce module de séparation sur les enregistrements DIADEMS a été aussi évaluée par les ethnomusicologues : un taux moyen d'efficacité de 90% a été obtenu. Le module de détection du chant est en cours d'évaluation.

Les quatrième et cinquième étapes sont réalisées en appliquant les méthodes développées lors de cette thèse sur le sous-corpus DIADEMS spécifique à la détection des tours de chant. Concernant la segmentation en tours de chant, une performance de 61,4% en termes de F-mesure a été obtenue avec notre système utilisant une sélection de bandes fréquentielles pertinentes pour la paramétrisation, une segmentation par BIC avec une fenêtre d'analyse de taille dynamique et une méthode de fusion par DCAP. Cette performance est plus faible que celle trouvée sur le corpus « studio » : ceci s'explique par l'hétérogénéité des enregistrements ethnomusicologiques. En effet, certains fichiers qui présentent des instruments percussifs ou un fort bruit de fond (principalement des cris) engendrent de nombreuses fausses alarmes. Concernant le regroupement en chanteurs, nous avons appliqué nos deux systèmes RDCAP et RDCAP+VCE. Le premier semble plus robuste que le second sur les enregistrements DIADEMS. En effet, en appliquant le premier système (constitué du regroupement par BIC et d'une fusion par RDCAP), nous obtenons un DER de 41,3% alors qu'en appliquant le second (possédant des étapes supplémentaires), le DER est plus élevé (48,4%). Des expériences complémentaires sur le système RDCAP ont permis d'améliorer le résultat assez nettement : baisse du DER de 5%.

Chapitre 7 :

Conclusion et perspectives

7.1. Conclusion

Dans ce travail de thèse, j'ai abordé le problème de la segmentation en tours de chant et celui de leur regroupement dans des enregistrements musicaux, ce qui m'a conduit à développer un système complet de segmentation et regroupement en chanteurs. Le contexte applicatif du projet DIADEMS en indexation a permis d'évaluer mon travail sur des données ethnomusicologiques allant au delà d'une classique évaluation réalisée sur un corpus d'enregistrements studio. Mes principales contributions ont été : la définition de la notion de « tour de chant » et la proposition de règles d'annotation manuelle associées, l'adaptation de méthodes standard de segmentation et regroupement en locuteurs dans le contexte du chant, la proposition d'une méthode de décision *a posteriori* pour pallier au problème de variabilité des paramètres intrinsèques de la méthode et la proposition d'une méthode de paramétrisation adaptée à notre tâche.

7.1.1. Vers une segmentation dynamique en tours de chant

Un des objectifs de cette thèse étant la réalisation de la segmentation en des zones acoustiquement homogènes par groupe de chanteur(s), une grande partie de mes travaux a été consacrée à l'adaptation des approches de segmentation existantes en parole au contexte du chant.

Une première méthode a été mise en œuvre à partir du Critère d'Information Bayésien (BIC) en utilisant les paramètres classiques de type MFCC pour la paramétrisation acoustique du chant. Cette première approche a permis de mettre en évidence les limites dans le contexte du chant en révélant deux problèmes. Le premier concerne l'influence de la taille de la fenêtre d'analyse dans laquelle un point de changement potentiel est recherché. Le deuxième est l'impossibilité de pouvoir fixer durant une phase de développement une unique valeur optimale du facteur de pénalité du BIC ; ces deux paramètres sont des éléments cruciaux pour la détection d'une frontière de segment.

Pour pallier au premier problème, nous avons implémenté une variante de la segmentation par BIC en s'inspirant des travaux de (Cettolo, et al., 2005), qui utilise une taille de fenêtre d'analyse dynamique, qui augmente tant qu'aucune frontière de segment n'est détectée. Ensuite, pour résoudre le problème de variabilité du facteur de pénalité λ, nous avons proposé une méthode de fusion (DCAP) qui permet de décider de l'existence d'une

frontière par consolidation, sous condition. Une frontière candidate n'est retenue que lorsqu'elle a été détectée par plusieurs systèmes de segmentation, qui utilisent des valeurs de λ différentes.

L'intérêt de notre méthode de segmentation réside dans sa robustesse vis-à-vis du contenu des enregistrements et des types de documents. De plus, le choix de la taille de la fenêtre d'analyse est fait de façon dynamique et le choix de la valeur du facteur de pénalité ne nécessite aucun ajustement *a priori*. Ces travaux ont donné 3 publications (Thlithi, et al., 2014), (Thlithi, et al., 2014a) et (Thlithi, et al., 2015).

7.1.2. Revisite de la paramétrisation

Au niveau caractérisation acoustique, nous avons testé plusieurs paramètres classiques utilisés en traitement de la parole comme les MFCC, les PLP et les FBANK, ainsi que des paramètres musicaux (chromas). Comme les FBANK étaient les paramètres donnant les meilleurs résultats, nous avons exploré leurs caractéristiques en testant plusieurs configurations et stratégies. Nous avons ainsi proposé une méthode de paramétrisation adaptée à notre contexte de segmentation en tours de chant qui consiste à utiliser seulement les 12 premières bandes parmi les 24 et à appliquer une stratégie de sélection de bandes fréquentielles pertinentes en conservant les bandes qui possèdent une variance importante, apparaissant comme les bandes les plus informatives.

Notre méthode de paramétrisation a permis d'avoir une performance satisfaisante sur le corpus de travail « studio » : une F-mesure de 75,8% avec les FBANK sélectionnés contre une F-mesure de 45,2% et de 69,1% avec 12 MFCC et 24 FBANK respectivement. Ce travail fait l'objet d'une soumission acceptée à CBMI'16 (Thlithi, et al., 2016).

7.1.3. Regroupement des groupes de chanteurs

Afin de regrouper les segments de chant d'un même groupe de chanteur(s) dans une seule classe, nous avons utilisé le BIC comme critère de décision de regroupement. Tout comme en segmentation, nos observations en regroupement en chanteurs ont montré qu'une unique valeur optimale du facteur de pénalité du BIC n'est pas adaptée au contexte de chant et nous avons proposé une méthode de fusion de plusieurs systèmes de regroupement (RDCAP) : un segment de chant est classé dans la classe qui a été choisie par la majorité des systèmes. Deux systèmes ont été proposés pour réaliser cet objectif :

- Le système RDCAP : celui-ci est composé d'une première passe de regroupement par BIC, sachant que le BIC que nous utilisons pour l'étape de regroupement est celui du LIUM, suivie d'une deuxième passe de fusion par RDCAP. L'application de ce système sur le corpus « studio » a donné un taux d'erreurs (DER) de 19,7%.

- Le système RDCAP+VCE : la première partie de ce système correspond à l'application du système RDCAP, et la deuxième partie est composée d'une re-segmentation par Viterbi suivie d'une deuxième passe de regroupement par le critère Entropie Croisée, sachant que le Viterbi et le CE que nous utilisons sont ceux du

LIUM. Ce système a permis d'avoir un DER de 10,9%, soit une baisse de 8,8% et 9,6% du taux d'erreurs par rapport au système RDCAP et au système initial du LIUM.

7.1.4. Application et mise en œuvre au sein de DIADEMS

Le contexte applicatif propre au projet DIADEMS m'a permis de tester mes méthodes sur des données très particulières qui sont des enregistrements ethnomusicologiques. La variété de ces données nous a imposé de mettre en place une chaîne de traitement complète afin d'aboutir à un résultat satisfaisant de segmentation et regroupement en chanteurs sur le corpus DIADEMS.

L'hétérogénéité de ce corpus en termes de contenu (bruit, parole, chant, instruments, etc.) nécessite un prétraitement pour détecter les zones de chant qui seront, par la suite, à l'entrée de notre système de segmentation et regroupement. La première étape de prétraitement consiste à repérer les zones d'intérêt en détectant les bruits techniques de démarrage de session. Nous avons proposé deux méthodes : la première consiste à détecter la décroissance de toutes les fréquences en forme de « 1/x », et la deuxième détecte le silence spécifique qui précède le phénomène « 1/x ». Les résultats de ces deux méthodes étaient satisfaisants avec une meilleure performance obtenue avec la deuxième approche. La deuxième étape réalise une détection des régions de musique en utilisant une méthode classique de l'équipe SAMoVA basée sur le « nombre de segments stables » et leur « durée » ; sa performance est jugée acceptable d'après une évaluation des ethnomusicologues partenaires du projet (resp. 68% et 53% de taux d'efficacité). L'étape finale de prétraitement effectue une détection de chant dans les zones de musique trouvées. L'outil de détection de chant utilisé tient compte du contexte polyphonique en réalisant une séparation monophonie / polyphonie avant d'entamer la détection de chant. L'application du module de séparation monophonie / polyphonie sur les données DIADEMS montre des résultats très satisfaisants (taux d'efficacité de 90%), sachant que la détection du chant est en cours d'évaluation par les ethnomusicologues.

L'application de notre système de segmentation sur le sous-corpus DIADEMS spécifique à la détection des tours de chant a donné une performance satisfaisante : F-mesure de 61,4%. L'application de nos systèmes de regroupement en chanteurs sur ce sous-corpus a montré que le système RDCAP est plus robuste que le système RDCAP+VCE sur les données DIADEMS : un DER de 41,3% est obtenu avec le système RDCAP contre un DER de 48,4% avec le système RDCAP+VCE. Une amélioration du système RDCAP a été proposée en adaptant l'intervalle de variation du facteur de pénalité du BIC aux données DIADEMS, ce qui a permis de réduire le taux d'erreurs pour atteindre un DER de 36,4%.

Ces performances sont prometteuses et mettent également en évidence les situations qui peuvent poser problème. La présence d'évènements sonores en fond (instruments percussifs, frappements de mains, cloches, bruit d'animaux, etc.) est assez problématique pour nos processus de segmentation et de regroupement en chanteurs.

7.2. Perspectives

Les travaux effectués lors de ce doctorat ont permis d'entamer un nouveau sujet dans le contexte du traitement du chant en posant la question « **qui chante et quand ?** ». Les résultats que nous avons obtenus sur un petit corpus de type studio sont satisfaisants, mais ils sont à confirmer par un passage à l'échelle (sur un corpus plus grand).

Cependant, les résultats que nous avons obtenus sont encore loin des performances de la même tâche en parole (segmentation et regroupement en locuteurs). Des améliorations semblent donc être possibles sur notre approche de segmentation en tours de chant. En effet, nous obtenons une performance de 75,8% sur le corpus « studio » alors qu'en segmentation en tours de parole les performances obtenues sont d'environ 90% avec les systèmes les plus performants. Notre approche de segmentation a montré que la détection des points de changement entre chanteurs est plus difficile quand il s'agit des situations de transition soliste-soliste et chœur-chœur que dans la situation de transition soliste-chœur et vice versa. Je pense que ce problème pourrait être résolu par une amélioration de la méthode de paramétrisation pour bien caractériser les voix des chanteurs et ainsi arriver à les différencier pour avoir une performance de détection similaire dans toutes les situations. Cela pourrait être réalisé en utilisant d'autres paramètres avec nos coefficients FBANK sélectionnés tels que le timbre qui permet de différencier deux sons de même hauteur, puissance et durée, et le vibrato puisque le rythme des oscillations sur la valeur de la fréquence fondamentale pour un chanteur donné est constant.

Un renforcement de la méthode de fusion pourrait aussi améliorer la performance de notre système de segmentation. Par exemple, nous pouvons remplacer le vote majoritaire simple par une combinaison plus intelligente par pondération des scores BIC obtenus avec différentes valeurs du facteur de pénalité. Nous pouvons estimer les facteurs de pondération sur un corpus de développement.

Concernant la mise en œuvre de notre système de segmentation et regroupement en chanteurs sur les données du projet DIADEMS, nous l'avons appliqué seulement sur un sous-corpus ne contenant que du chant et nous avons réalisé quelques adaptations afin d'améliorer les résultats sur ces données. Néanmoins, la performance reste toujours plus faible que celle obtenue sur le corpus « studio » à cause principalement de la présence d'autres événements sonores en fond. Ce problème pourrait peut-être être résolu en ajoutant un module de séparation de sources afin d'isoler le chant et de n'appliquer, par la suite, notre système de segmentation et regroupement que sur le chant. En effet, la séparation de sources permet de retrouver un signal source ou un ensemble de signaux sources à partir d'une ou plusieurs observations de leur mélange en tenant compte de la structure de mélange et en faisant des hypothèses sur les sources. La séparation de sources a été utilisée dans des applications d'indexation automatique de documents multimédia contenant plusieurs objets sonores (Parvaix, 2010). Donc, son application pourrait permettre de retrouver les sources de la musique vocale (chant) parmi les autres sources (instruments, bruits…).

De plus, une constitution d'un corpus « studio » avec chant et instruments pourrait être utile afin de comparer les performances qui seront obtenues sur ce nouveau corpus avec le corpus DIADEMS qui contient des instruments et aussi pour étudier la robustesse de notre système par rapport aux contenus des enregistrements musicaux.

A plus long terme, il serait intéressant d'étudier la possibilité de généraliser notre approche afin de réaliser une segmentation en tours de musique dont le principe est illustré dans la Figure 7. 1. En effet, grâce à la méthodologie utilisée, fondée sur le BIC, il semble raisonnable de penser que ce soit possible. Pour cela, il conviendrait de réaliser une annotation manuelle difficile, tenant compte des entrées-sorties de chaque source, c'est-à-dire chaque instrument et chaque chanteur.

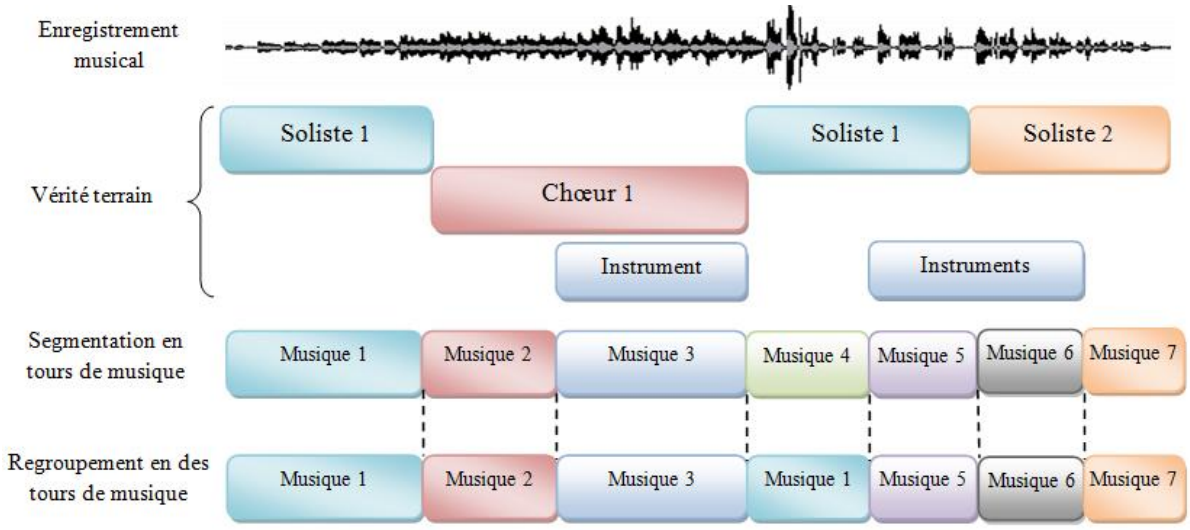

Figure 7. 1 – Illustration du principe de segmentation et regroupement en des tours de musique.

Les résultats de la segmentation en tours de musique pourraient ainsi servir de prétraitement pour d'autres applications tell que l'identification du nombre de sources.

Bibliographie

Abad, A, Rodriguez-Fuentes, L.J et Penagarikano, M. 2013. On the Calibration and Fusion of Heterogeneous Spoken Term Detection Systems. *INTERSPEECH.* 2013.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic and Control*. 1974, pp. 716-723.

Ajmera, J, Bourlard, H, Lapidot, I et Mccowan, I. 2002. Unknown-multiple speaker clustering using hmm. *ICSLP-2002.* 2002, pp. 573-576.

André-Obrecht, R. 1988. A New Statistical Approach for Automatic Speech Segmentation. *IEEE Transactions on Audio, Speech, and Signal Processing.* Janvier 1988, pp. 29-40.

André-Obrecht, R. 1998. A new statistical approach for the Automatic Segmentation of Continuous Speech Signals. *IEEE Transactions on Audio, Speech, and Language Processing.* 1998, Vol. 36-1, pp. 29-40.

Anguera Miro, X. 2006. Robust Speaker Diarization For Meetings. *Thèse*. 2006.

ANSI. 1960. ANSI American National Standards Institute. [En ligne] 1960. http ://www.ansi.org.

Berenzweig, A et Ellis, D. 2001. Locating Singing Voice Segments Within Music Signals. *Workshop on the applications of signal processing to audio and acoustics.* 2001, pp. 119-122.

Boite, R, Bourlard, H, Dutoit, T, Hancq, J et Leich, H. 2000.*Traitement de la parole.* 2000.

Bost, X et Linarès, G. 2014. Constrained speaker diarization of TV series based on visual patterns. *IEEE/ISCA Speech And Language Technology Workshop, SLT*. 2014, South Lake Tao, Nevada, USA.

Bost, X, Linarès, G et Gueye, S. 2015. Audiovisual speaker diarization of TV series. *IEEE Intenational Conference on Audio, Speech and Signal Processing, ICASSP*. 2015, Brisbane Australia.

Bousquet, P-M, Matrouf, D et Bonastre, J-F. 2011. Intersession compensation and scoring methods in the i-vectors space for speaker recognition. *Interspeech'11.* 2011.

Calliope. 1989. La parole et son traitement automatique. 1989.

Castellengo, M. 2007. Compte rendu des recherches 2002-2007. *Présentation des recherches*. 2007.

Cettolo, M, Vescovi, M et Rizzi, R. 2005. Evaluation of BIC-based algorithms for audio segmentation. *Computer Speech And Language.* 2005, pp. 147-170.

Chen, S et Gopalakrishnan, P.S. 1998. Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion. *DARPA Broadcast News Transcription and Understanding Workshop.* 1998, pp. 127-132.

Chen, Scott S et Gopalakrishnan, P. S. 1998. Clustering via the bayesian information criterion with applications in speech recognition. 1998, Vol. 2, pp. 645-648.

Cho, Y. D, Kim, M.Y et Kim, S.R. 1998. A Spectrally Mixed Exitation (SMX) Vocoder with Robust Parameter Determination. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* . 1998.

Chou, W et Gu, L. 2001. Robust singing detection in speech/music discriminator design. *International Conference on Acoustics, Speech, and Signal Processing.* 2001, pp. 865-868.

Chuan Toh, C, Zhang, B et Wang, Y. 2008. Multiple-Feature Fusion Based Onset Detection for Solo Singing Voice. *International Conference on Music Information Retrieval (ISMIR).* 2008, pp. 515-520.

Daniel, P et Ellis, W. 2007. Classifying music audio with timbral and chroma features. *Austrian Computer Society.* 2007.

DGA, Délégation Général de l'Armement. 2008. Convention d'annotation détaillée et enrichie. *ESTER 2.* 2008.

De Cheveigné, A et KAWAHARA, H. 2002. YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America.* 2002, Vol. 111(4), pp. 1917-1930.

Dehak, N, Kenny, P, Dehak, R, Dumouchel, P et Ouellet, P. 2011. Front-end factor analysis for speaker verification. *IEEE TASLP.* 2011, Vol. 19, pp. 788-798.

Delacourt, P, Kryze, D et Wellekens, C. 1999. Speaker-based segmentation for audio data indexing. *ESCA ETRW Workshop Accessing Information in Spoken Audio.* 1999.

Delacourt, P et Wellekens, C. J. 2000. Distbic: a speaker-based segmentation for audio data indexing. *Speech Commun.* 2000, Vol. 32(1-2), pp. 111-126.

Dellaert, F. 2002. The Expectation Maximization Algorithm. *Rapport technique de l'institut technique de Georgia.* 2002.

El-Khoury, E, Senac, C et Pinquier, J. 2009. Improved speaker diarization system for meetings. *ICASSP.* 2009, pp. 4097-4100.

El-Khoury, E. 2010. Unsupervised video indexing based on audiovisual characterization of persons. *Thèse.* Toulouse. 2010.

Ester2. 2008. Plan d'évaluation Ester 2 phases 1 & 2. *Evaluation des systèmes de transcription enrichie d'émissions radiophoniques.*

Ezzaidi, H et Rouat, J. 2002. Speech, Music and Songs Discrimination in the Context of Handsets Variability. *ICSLP.* 2002, pp. 16-20.

Foote, J. 2000. Automatic audio segmentation using a measure of audio novelty. *Proceedings of the IEEE International Conference on Multimedia and Expo.* 2000.

Foote, J et Uchihashi, S. 2001. The Beat Spectrum : a New Approach to Rhythm Analysis. *IEEE International Conference on Multimedia and Expo (ICME).* 2001, pp. 881-884.

Fujishima, T. 1999. Realtime chord recognition of musical sound : A system using common lisp music. *International Computer Music Conference (ICMC).* 1999.

Galliano, S, Gravier, G et Chaubard, L. 2009. The ester 2 evaluation campaign for the rich transcription of French radio broadcasts. *Interspeech'2009.* 2009.

Gauvain, J.L et Lee, C.H. 1994. Maximum *a Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Transactions on Speech and Audio Processing.* 1994, Vol. 2, pp. 291-299.

Gish, H, Siu, M.H et Rohlicek., R. 1991. Segregation of speakers for speech recognition and speaker identification. *ICASSP Proceedings of the Acoustics, Speech, and Signal Processing.* 1991, pp. 873-876.

Gomez, E. 2006. Tonal Description of Polyphonic Audio for Music Content Processing. *INFORMS Journal on Computing.* 2006, pp. 294-304.

Gravier, G, Adda, G, Paulsson, N, Carré, M, Giraudel, A et Galibert, O. ETAPE : The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. *Association Francophone de la Communication Parlée (AFCP).* 2012.

Henrich Bernardoni, N. 2014. Notre voix, cet instrument de tous les possibles **.** *Palais des Congrès, Le Mans, Rencontre organisée par Maine Sciences dans le cadre de la XXXè édition des Journées d'Études sur la Parole et de la biennale Le Mans Acoustique 2014.*

Henrich Bernardoni, N. 2014. Du souffle à la parole et au chant : notre voix s'exprime . *Science et Musique sur une même partition, dans le cadre des « Fondamentales du CNRS ». 2014.*

Hermansky, H, Morgan, N, Bayya, A et Kohn, P. 1991. RASTA-PLP speech analysis. *International Computer Science Institute.* 1991.

Houtgast, T et Steeneken, J. M. 1985. A Review of the MTF Concept in Room Acoustics and its Use for Estimating Speech Intelligibility in Auditoria. *Journal of the Acoustical.* 1985, Vol. 77, pp. 1069-1077.

Izmirli, O. 2005. Template Based Key Finding from Audio. *International Computer Music Conference (ICMC).* 2005, pp. 211-214.

Kaiser, F et Peeters, G. 2014. Adaptive Temporal Modeling of Audio Features in the Context of Music Structure Segmentation. *Springer International Publishing Switzerland.* 2014, pp. 248-261.

Klapuri, A. 2006. Multiple Fundamental Frequency Estimation by Summing Harmonic Amplitudes. *International Conference on Music Information Retrieval (ISMIR).* 2006.

Kullback, S et Leibler, R. A. 1951. On Information and Sufficiency**.** *The Annals of Mathematical Statistics.* 1951, pp. 79-86.

Lachambre, H. 2008. Estimation of Weibull bivariate distribution parameters via the moment method. *Rapport technique.* 2008.

Lachambre, H, André-Obrecht, R et Pinquier, J. 2009. Caractérisation de la voix chantée en contexte monophonique et polyphonique. *Groupe de Recherche et d'Etudes du Traitement du Signal et des Images (GRETSI).* 2009.

Lachambre, H. 2009. Caractérisation de l'environnement musical dans les documents audivisuels. *Thèse*. 2009.

Lachambre, H, André-Obrecht, R et Pinquier, J. 2009. Singing voice detection in monophonic and polyphonic contexts. *European Signal Processing Conference.* 2009, pp. 1344-1348.

Lapidot, I. 2003. Som as likelihood estimator for speaker clustering. *Eurospeech-2003.* 2003, pp. 3001-3004.

Le Coz, M, André-Obrecht, R et Pinquier, J. 2012. Feasibility of the Detection of Choirs for Ethnomusicologic Music Indexing. *International Workshop on Content-Based Multimedia Indexing.* 2012, pp. 145-148.

Le Coz, M, Lachambre, H, Koenig, L et André-Obrecht, R. 2010. A segmentation-based tempo induction method. *International Society for Music Information Retrieval Conference.* 2010, pp. 27-31.

Le Coz, Maxime, Doctorat. 2014. *Spectre de rythme et sources multiples au coeur des contenus ethnomusicologiques et sonores.* 2014.

Le, V. B, Mella, O et Fohr, D. 2007. Speaker diarization using normalized cross-likelihood ratio. *International Conference on Spoken Language Processing (ISCA).* 2007.

Lu, L, Jiang, H et Zhang, H. 2001. A Robust Audio Classification and Segmentation. *ACM International Conference on Multimedia.* Septembre 2001, pp. 203-211.

Lukashevich, H, Gruhne, M et Dittmar, C. 2007. Effective Singing Voice Detection in Popular Music using ARMA Filtering. *International Conference on Digital Audio Effects.* 2007.

Markaki, M, Holzapfel, A et Stylianou, Y. 2008. Singing Voice Detection using Modulation Frequency Features. *Workshop on Statistical And Perceptual Audition (SAPA).* 2008.

Marozeau, J. 2004. L'effet de la fréquence fondamentale sur le timbre. *Thèse.* 2004.

Meignier, S, Bonastre, J-F et Igounet, S. 2001. E-hmm approach for learning and adapting sound models for speaker indexing. *Odyssey Speaker and Language Recognition Workshop.* 2001, pp. 175-180.

Meignier, S et Merlin, T. 2009. LIUM SpkDiarization: an open-source toolkit for diarization. *CMU SPUD Workshop.* 2009.

Meignier, S. 2015. Détection et identification des locuteurs des émissions radiophoniques et télévisées. *HDR.* 2015.

Meron, Y et Hirose, K. 2000. Synthesis of Vibrato Singing. *International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 2000, Vol. 2, pp. 745-748.

Mesaros, A et Moldovan, C. 2006. Method for Singing Voice Identification using Energy Coefficients as Features. *IEEE International Conference on Automation, Quality and Testing, Robotics.* 2006, pp. 161-166.

NIST. 2003. The rich transcription spring 2003 (rt-03s) evaluation plan. 2003.

Parvaix, M. 2010. Séparation de sources audio informée par tatouage pour mélanges linéaires instantanés stationnaires. *Thèse.* 2010.

Peeters, G. 2005. Rhythm Classification using Spectral Rhythm Patterns. *International Conference on Music Information Retrieval (ISMIR).* 2005.

Peeters, G. 2007. Template-Based Estimation of Time-Varying Tempo. *EURASIP Journal on Advances in Signal Processing.* 2007.

Pinquier, J. 2004. Indexation sonore : recherche de composantes primaires pour une structuration audiovisuelle. *Thèse.* 2004.

Pinquier, J, Rouas, J-L et André-Obrecht, R. 2002. Robust speech/music classification in audio documents. *Seventh International Conference on Spoken Language Processing.* 2002.

Pols, J. C. W, Van Der Kamp, L.J et Plomp, R. 1969. Perceptual and physical space of vowel sounds. *Journal of Acoustical Society of America.* 1969, pp. 458-467.

Purwins, H. 2005. Profiles of Pitch Classes — Circularity of Relative Pitch and Key : Experiments, Models, Music Analysis, and Perspectives. *Thèse.* 2005.

Ramona, M, Richard, G et David, B. 2008. Vocal Detection in Music with Support Vector Machines. *International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 2008, pp. 1885-1888.

Regnier, L et Peeters, G. 2009. Singing Voice Detection in Music Tracks using Direct Voice Vibrato Detection. *International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 2009, pp. 1685-1688.

Reynolds, D. A, Singer, E, Carlson, B. A, OOLeary, G.C, McLaughlin, J.J et Zissman, M. A. 1998. Blind clustering of speech utterances based on speaker and language characteristics. *International Conference on Spoken Language Processing.* 1998.

Rissanen, J. 1989. Stochastic Complexity in Statistical Inquiry Theory. *World Scientific Publishing.* 1989.

Rocamora, M et Herrera, P. 2007. Comparing Audio Descriptors for Singing Voice Detection in Music Audio Files. *Brazilian Symposium on Computer Music.* 2007.

Rossignol, S, Depalle, P, Soumagne, J, Rodet, X et Collette, J-L. 1999. Vibrato : Detection, Estimation, Extraction, Modification. *International Conference on Digital Audio Effects.* 1999.

RT03. 2003. Guidelines for RT-03 Transcription. *Linguistic Data Consortium.* 2003.

RT09. 2009. The 2009 (RT-09) Rich Transcription. *Meeting Recognition Evaluation Plan.* 2009.

Saunders, J. 1996. Real-time Discrimination of Broadcast Speech/Music. *IEEE International Conference on Audio, Speech and Signal Processing.* Mai 1996, pp. pages 993–996.

Scheirer, E et Slaney, M. 1997. Construction and Evaluation of a Robust Multifeature. *IEEE International Conference on Audio, Speech.* Avril 1997, pp. 1331-1334.

Scheirer, E. 1997. Pulse Tracking with a Pitch Tracker. *IEEE Workshop on Application of signals Processing to Audio and Acoustics.* 1997.

Seashore, C.E. 1938. Psychology of Music. *McGraw-Hill Book Company.* 1938.

Shen, J, et al. 2006. Towards efficient automated singer identification in large music. *The 29th Annual International ACM SIGIR conference on Research and Development in Information Retrieval.* 2006, pp. 59-66.

Shepard, R. 1964. Circularity in Judgments of Relative Pitch. *Journal of the Acoustical Society of America.* 1964.

Schwarz, G. 1978. Estimating the dimension of a model. *The annals of Statistic.* 1978, Vol. 6, pp. 461-464.

Siegler, M.A, Jain, U, Raj, B et Stern, R.M. 1997. Automatic segmentation, classification and clustering of broadcast news audio. *DARPA Speech Recognition Workshop.* 1997, pp. 97-99.

Sivakumaran, P, Fortuna, J et Ariyaeeinia, A.M. 2001. On the use of the bayesian information criterion in multiple speaker detection. *The 7th European Conference on Speech Communication and Technology (Eurospeech'01).* 2001, pp. 795-798.

Solomonoff, A, Mielke, A, Schmidt, M et Gish, H. 1998. Clustering speakers by their voices. *International Conference on Acoustics Speech and Signal Processing (IEEE, ICASSP 98).* Mai 1998.

Sundberg, J. 1994. Acoustic and Psychoacoustic Aspects of Vocal Vibrato. *STL-QPSR.* 1994, pp. 45-68.

Taniguchi, Toru, et al. 2005. Discrimination of Speech, Musical Istruments and Singing Voices Using the Temporal Patterns of Sinusoidal Segments in Audio Signals. *Interspeech - European Conference on Speech Communication and Technology. ISCA.* 2005.

Thlithi, M, Pellegrini, T, Pinquier, J, André-Obrecht, R, Guyot, P. 2014. Segmentation in singer turns with the Bayesian Information Criterion. *INTERSPEECH.* 2014.

Thlithi, M, Pellegrini, T, Pinquier, J, André-Obrecht, R, Guyot, P . et al. 2014. Application du critère BIC pour la segmentation en tours de chant. *Journées d'Etudes sur la Parole.* 2014.

Thlithi, M, Barras, C, Pinquier, J, Pellegrini, T. 2015. Singer diarization: application to ethnomusicological recordings. *Folk Music Analysis (FMA)*124-125. 2015.

Thlithi, M, Pinquier, J, Pellegrini, T et André-Obrecht, R. 2014. FilterBank coefficients selection for segmentation in singer turns. *International Workshop on Content-based Multimedia Indexing (CBMI).* 2016.

Timmers, R et Desain, P. 2000. Vibrato : Questions and Answers from Musicians and Science. *International Conference on Music Perception and Cognition.* 2000.

Tsai, W. H et Wang, H. M. 2007. Speaker clustering based on minimum rand index. *International Conference on Acoustics, Speech, and Signal Processing.* 2007, pp. 485-488.

Tsai, W. H, et al. 2005. Clustering speech utterances by speaker using eigenvoice-motivated vector space model. *International Conference on Acoustics, Speech, and Signal Processing.* 2005, pp. 725-728.

UCL. 2002. Tours de parole. *http://www.fltr.ucl.ac.be/FLTR/ROM/FOREO/tourparole/negocier_theorie.swf.* [En ligne] Faculté de philosohphie arts et lettres, 2002.

Wagsta, K, Cardie, C, Rogers, S et Schroedl, S. 2001. Constrained K-means Clustering with Background Knowledge. *International Conference on Machine Learning.* 2001, pp. 577-584.

Wold, E, et al. 1999. Classification, Search and Retrieval of Audio. *CRC Handbook of multimedia computing.* 1999.

Yu, Y, Crucianu, M, Oria, V et Damiani, E. 2010. Combining multi-probe histogram and order-statistics based lsh for scalable audio content retrieval. *ACM Multimedia.* 2010.

Zhang, T. 2002. System and method for automatic singer identification. *HP Labs Technical Report.* 2002.

Zhang, T. 2003. Automatic Singer Identification. *IEEE International Conference on Multimedia and Expo (ICME).* 2003, pp. 33-36.

Zhang, T, Kuo, C et J, C. 1998. Hierarchical System for Content-Based Audio Classification. *Conference on Multimedia Storage and Archiving Systems III.* 1998, pp. 398-409.

Zhu, X, Barras, C, Meignier, S et Gauvain, J.L. 2005. Combining speaker identification and bic for speaker diarization. *Interspeech.* 2005, pp. 2441-2444.